DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY


Three Essays on Data-Driven Personalization and Targeting for Marketing Interventions

Citation

Huang, Ta-Wei. 2025. Three Essays on Data-Driven Personalization and Targeting for Marketing Interventions. Doctoral Dissertation, Harvard University Graduate School of Arts and Sciences.

Link

https://dash.harvard.edu/handle/1/42719708

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFlOWI2M2EwYTg

Accessibility

https://accessibility.huit.harvard.edu/digital-accessibility-policy

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story

HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the Committee for the PhD in Business Administration have examined a dissertation entitled

Three Essays on Data-Driven Personalization and Targeting for Marketing Interventions

Presented by Ta-Wei Huang

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature _____
Eva Ascarza, Chair

Signature _____
Sunil Gupta

Signature _____
Ayelet Israeli

Signature _____
Jeremy Yang

Signature _____
José Zubizarreta

Date: April 14, 2025

Three Essays on Data-Driven Personalization and Targeting for Marketing Interventions

A dissertation presented

by

Ta-Wei Huang

to

The Department of Marketing at the Harvard Business School

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Business Administration

Harvard University

Cambridge, Massachusetts

April 2025

© 2025 Ta-Wei Huang

All rights reserved.

Dissertation Advisor: Eva Ascarza Ta-Wei Huang

Three Essays on Data-Driven Personalization and Targeting for Marketing Interventions

Abstract

Companies today face significant challenges in personalizing marketing interventions while balancing long-term objectives, adhering to privacy regulations, and managing complex decision spaces. This dissertation addresses these core issues through three interconnected essays and provides practical methodologies for enhancing personalized marketing interventions in today's data-driven environment.

The first essay (Chapter 1) demonstrates and tackles the challenges of optimizing long-term business performance through targeted interventions. It highlights how cumulative unexplained variations in repeated customer behavior can weaken the direct optimization of long-term outcomes. To address this, the essay introduces a surrogate index methodology using short-term signals, coupled with a novel separate imputation strategy to handle the churn and purchase processes driving customer value. Through simulations and a real-world marketing application, the essay demonstrates that the proposed approach significantly outperforms traditional methods in enhancing long-term targeting effectiveness.

The second essay (Chapter 2) examines how *Local Differential Privacy* (LDP) — a strong privacy technique that introduces noise into individual-level data — affects the accuracy of personalized marketing interventions based on *Conditional Average Treatment Effect* (CATE) predictions. We show that LDP induces heterogeneous and model-dependent errors that hinder accurate personalization. To address these limitations, we propose an honest post-processing method using an unbiased but noisy proxy combined with iterative boosting and a subgroup cross-learning strategy to ensure honesty and mitigate overfitting. Empirical tests demonstrate that our method significantly improves prediction accuracy and treatment prioritization, enabling organizations to achieve effective personalization despite privacy constraints.

The final essay (Chapter 3) introduces *Incrementality Representation Learning* (IRL), a novel multitask framework for predicting heterogeneous causal effects of marketing interventions. By leveraging past experiments, IRL efficiently designs and targets personalized interventions without extensive testing. It extracts generalizable low-dimensional representations of intervention features and customer covariates. Empirical validation using data from 274 promotional campaigns demonstrates that IRL significantly enhances targeting accuracy and effectively generalizes predictions to both known and untested interventions and customer segments, addressing challenges in high-dimensional decision spaces and cold-start scenarios. Additionally, the essay develops a decision framework and interpretation tool to assist firms in identifying critical design features and tailoring promotions for maximum profitability.

Contents

List of Tables

List of Figures

Acknowledgments

For this fruit, I impose no condition, but may you now learn about ambition.

— Robert Fisher, *The Knight in Rusty Armor*.

I owe my journey thus far to the guidance, support, and love of many. Words cannot fully express the depth of my gratitude toward these individuals. I will forever be indebted to them, and I commit myself to continuously contributing to knowledge and society to give back the support I have received.

To my committee — Eva Ascarza, Ayelet Israeli, Sunil Gupta, Jeremy Yang, and José R. Zubizarreta — thank you for believing in me and for shaping me into the scholar I am today. Eva, your mentorship has meant the world to me. You've taught me how to think deeply, aim high, and stay grounded in kindness. I look up to you not just as a scholar but as a person. Ayelet, your sharp insights and honest feedback have pushed my work to be better at every stage, and your support — both academic and personal — has helped me through difficult moments. Sunil, your perspective has always reminded me to think bigger and consider the impact of my work beyond the page. Jeremy, your constant encouragement and thoughtful advice have been a source of motivation. José, you broadened both the breadth and depth of my understanding in the field and have profoundly shaped the way I approach research. I'm deeply grateful to each of you — for your guidance, your belief in me, and the time and care you've invested in my growth.

I am sincerely thankful for my community at HBS, whose warmth and camaraderie I will truly miss. To the HBS Marketing department — it's been a privilege to be part of such a supportive group. Special thanks to Elie Ofek and Shunyuan Zhang for your early guidance and mentorship, and to Jimin Nam and Lucy Shen for your unwavering support even beyond graduation. My gratitude also extends to the HBS Office of Doctoral Programs for providing invaluable resources and thoughtful care throughout my PhD journey. To my friends at HBS and members of the Customer Intelligence Lab, your friendship and shared experiences — through both challenges and celebrations — have profoundly enriched my life.

Outside academia, I feel incredibly lucky to have been surrounded by wonderful friends in Boston — Chi Kuan, Jerry Lin, Jean Lee, Tsung-Hao Ku, Alice Yang, Alban Yau, Jean Ma, Lois Tang,

Hsiang Hsu, Martin Söndergaard, and many others — who made this city truly feel like a second home. Your warmth, laughter, and presence have provided comfort and joy, turning ordinary moments into lasting memories. To my friends in Taiwan and around the world, your unwavering support from afar has lifted my spirits and kept me grounded throughout this journey.

Finally, to my parents — my foundation — I am endlessly grateful for your unconditional love, unwavering support, and the countless sacrifices you've made for me. Your faith in me, especially during times when I doubted myself, has given me the courage to pursue my dreams. I am proud to be your son, and I hope I've made you proud too.

This dissertation is dedicated to my parents, who have always inspired me to strive to be the best version of myself.

Chapter 1

Doing More with Less: Overcoming Ineffective Long-term Targeting Using Short-Term Signals[1]

1.1 Introduction

Recent advancements in business experimentation combined with machine learning have transformed how companies execute targeted interventions. Through controlled experiments, businesses can infer causal relationships between their marketing offerings and customers' responses. Rather than simply measuring the average impact across all customers, companies can further identify differences in customer sensitivity based on individual characteristics, commonly quantified as the conditional average treatment effect (CATE). This approach empowers firms to focus on customers predicted to respond most favorably to their objectives (e.g., profits or purchases), particularly those with the highest predicted CATEs. This method has gained significant popularity among organizations to design effective targeted intervention, and some tech companies, such as Microsoft [157] and Uber [36], have taken a step further by open-sourcing their tools for CATE estimation. This has enabled more companies to adopt this approach and

[1]Coauthored with Eva Ascarza. This chapter is based on a previously published article titled "Doing More with Less: Overcoming Ineffective Long-term Targeting Using Short-Term Signals," which appeared in *Marketing Science* in 2024.

develop highly precise targeted marketing interventions at scale.

This *test-to-target* approach has proven effective in various marketing contexts, such as customer retention [93, 9, 137], membership subscription [184, 218], and catalog mailing purchases [103]. Despite its popularity, it remains untested whether this approach can effectively optimize long-term outcomes, such as customer lifetime value (CLV) or repeated purchases, which are typically the top-line metrics for a firm. In theory, the observation window should not alter the way firms optimize their resource allocation — if the business goal is to maximize long-run outcomes, firms should target customers based on their *long-term sensitivity* to the intervention, measured as the CATE of the intervention on the long-term business outcome.

However, our research shows that the conventional test-to-target approach can be ineffective at optimizing long-term outcomes, especially those driven by *recurring customer behaviors*, such as total purchases over an extended post-intervention period. Unlike short-term outcomes (e.g., immediate purchases after the intervention), long-term outcomes accumulate individual customer behaviors, such as unobserved heterogeneity or unexplained customer attrition, that cannot be explained by the observed customer characteristics collected before the intervention. As a result, long-term outcomes not only carry information about the treatment effect (which is what CATE models aim to capture), but also accumulate unexplained variations that grow over time. This presents a significant and understudied challenge in CATE estimation, as existing models may generate unstable and high-variance CATE predictions when unexplained variations are large. Targeting customers based on such high-variance models can result in an ineffective targeting strategy when optimizing for long-term outcomes.

This paper has two main objectives. First, we examine the challenge of estimating CATEs for *long-term, recurrent customer behaviors*. Our theoretical analysis shows that such outcomes can accumulate unexplained variations due to *unobserved heterogeneity* or *customer attrition*. Consequently, as we extend the observation window, the outcome variance — and by extension, the variance of most CATE models — tends to rise. This increased variance inadvertently heightens the likelihood of incorrect targeting. Second, we present a solution that enables firms to implement more effective targeted interventions and achieve better long-term performance. We recommend that firms create a *noise-reduced proxy* based on immediate post-intervention behaviors and use this proxy for CATE estimation instead of the actual long-term outcome. Although it may seem

counterintuitive, this method can potentially minimize unexplained variations and effectively capture long-term treatment effects that are reflected in short-term behavioral changes.

To construct this noise-reduced proxy, we adopt the surrogate index approach [13, 212], which uses historical data that is readily available to firms to infer the relationship between short-term behaviors and the long-term outcome. In addition to providing valid CATE estimation, we formally show that the surrogate index has smaller unexplained variations than the actual long-term outcome, leading to more accurate CATE estimation. This enables firms to effectively target customers based on their long-term sensitivity to the intervention while mitigating the noise accumulation problem.

We emphasize that the conventional method of surrogate index construction, as recommended by [16] and [212], may not be optimal in scenarios with customer attrition. In the presence of attrition, long-term repeated behaviors become the product of two critical variables: the duration of a customer's lifetime and the intensity of these behaviors during that lifetime. As both these variables are influenced by short-term behaviors and idiosyncratic variations in each period, the conventional modeling approach fails to identify the correct relationship between short-term and long-term outcomes [e.g., 31, 106, 187].

To address this issue, we propose a novel *separate imputation* technique. Our approach involves developing two separate models using historical data: one to predict the time of a customer's last observed purchase (i.e., the proxy for lifetime) and another to predict the average purchases per period when a customer is still active. We then combine the predictions of both models to estimate expected future purchases. This approach stands out from other surrogate models by effectively mitigating the issue of increased variance that arises from the inseparable nature of customer attrition and purchase intensity. Consequently, this technique enables firms to construct more accurate and robust surrogate indices, leading to improved CATE estimation and more effective targeting strategies for optimizing long-term outcomes.

Through simulation analyses and a real-world marketing campaign, we demonstrate that our proposed approach is significantly more effective than relying on the actual long-term outcome. Specifically, we show that targeting rules based on immediate, short-term signals — either directly from short-term outcomes or through surrogate indices — consistently yield better results than those using CATE models estimated using the long-term outcome. This is surprising because

it suggests that, for optimal long-term performances, firms should rely on short-term outcomes and historical information rather than on the long-term outcomes themselves. Furthermore, we demonstrate that our separate imputation approach achieves the best targeting performance. In the real-world application, targeting customers using the proposed solution yields a 6% increase in profits (compared to directly rolling out the best action to all customers), while targeting based on the long-term outcome results in a 3% profit loss.

There are several compelling reasons for firms to implement our proposed solution. Firstly, it utilizes existing historical data, obviating the need for new experimental data to reduce variance. This benefit translates into cost savings, as firms can bypass the expenses associated with expanding experimental samples. Secondly, the solution integrates seamlessly with standard machine learning algorithms and existing software packages, facilitating a swift and effective deployment geared towards higher profits. Thirdly, it is applicable to a wide variety of business settings, including retailers, e-commerce, apparel, and non-profit organizations. Lastly, our method expedites decision-making processes. Firms no longer have to endure the typical delays associated with waiting for long-term outcomes, thereby accelerating the implementation of targeted interventions [13, 212].

Our research contributes to the literature in four strands. First, we address practical challenges in designing and implementing targeting policies. We underscore the limitations of current best practices [e.g., 9, 184, 63, 218], particularly in their ineffectiveness for optimizing noisy long-term outcomes. We contribute to this literature by proposing a new targeting paradigm where firms *reduce noise* in the outcome variable *before* estimating any CATE model. Although ignoring the actual outcome of interest may seem counterintuitive, we demonstrate how creating the "right" proxy using the surrogate index approach with proper imputation methods results in more effective targeting.

Second, our work highlights a significant challenge in estimating the CATE and effective targeting in scenarios characterized by a low signal-to-noise ratio. Considerable research has been dedicated to developing methods for CATE estimation [114, 113, 94, 90, 41, 134, 15, 126, 155] and policy learning [143, 133, 17, 147]. To the best of our knowledge, this paper is the first to theoretically explore how unexplained variations affect the predictive accuracy of CATE models and their targeting performance. Moreover, we incorporate behavioral insights from marketing

literature to illustrate why the issue of high noise is common in many marketing contexts. Our study provides important insights into the limitations of existing CATE models and highlights the need for robust solutions to estimate CATEs.

Third, we contribute to the literature on statistical surrogacy and long-term treatment effect estimation [164, 13, 212, 165, 115]. This literature has traditionally assumed that firms use short-term signals because of the cost of waiting to observe long-term outcomes. We further demonstrate, using formal theory and empirical evidence, the value of leveraging short-term proxies even when the actual long-term outcome is observed. Moreover, past research has primarily constructed surrogate indices using standard regression models, while our work highlights the importance of considering the data generating process for surrogate indices construction.

Lastly, our work contributes to the literature on treatment effect estimation for low-sensitivity experiments (i.e., experiments with outcome variance much larger than the treatment effect). Our approach differs from previous research [e.g., 53, 95, 120] in a critical way — we do not reduce variance by eliminating variations that can be explained by customer observables. Instead, we advocate for the use of information explainable by short-term signals and pre-treatment covariates for CATE estimation and targeting. This strategy is pivotal, as it targets the core issue of high variance in CATE estimates arising from unexplained variations, rather than from the observable heterogeneity among customers.

1.2 Data and Motivation

We use a field experiment from a retail-technology company in Taiwan to highlight the challenges of targeting with the goal of maximizing repeated purchases over an extended period of time. This company manages a network of self-service vending machines at various locations within a city. Customers can grab food and beverages from a vending machine, and the machine automatically counts the items using RFID technology and charges the customers through their pre-registered payment methods. The company uses a third-party messaging platform (like WhatsApp) to manage customer profiles and send marketing messages. To use this service, customers must join the company's messaging app channel and register their payment methods through the messaging app.

1.2.1 The Marketing Intervention

As part of the customer activation process, the company sends a 15% discount coupon to every new customer following their first purchase. The coupon is automatically applied to the next purchase made within 14 days, after which it expires. The company considered offering additional coupons to some newly acquired customers, but only if doing so would increase their total purchases in the subsequent months. To develop a targeted approach, they conducted a randomized controlled experiment to identify customers who would increase their purchases as a result of such an intervention.

In the experiment, the company randomly assigned customers who had just made their first purchase to one of two groups: a control group ($W_i = 0$) that received one coupon (i.e., the "business-as-usual" case) or a treatment group ($W_i = 1$) that received three coupons. All coupons offered a 15% discount and expired after 14 days. The experiment included 1,853 customers, with 889 assigned to the treatment group and 964 to the control group. The company collected several pre-treatment covariates on customers to design personalized interventions, such as acquisition channels and information about their first purchase. (Further details on these covariates and randomization checks can be found in Table 1.2 in Section 1.6.)

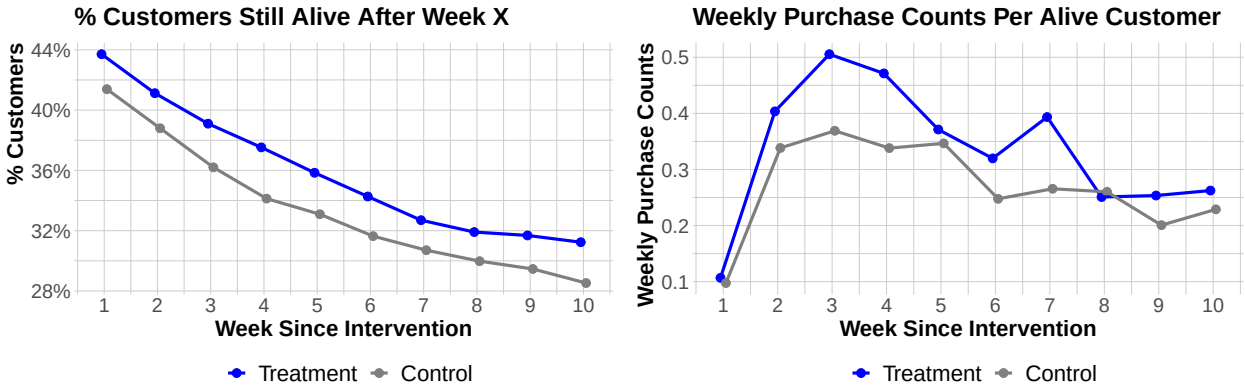
1.2.2 Average Treatment Effect on Repeated Purchasing Behaviors

Before delving into customer-specific impacts, we first assess the average treatment effect of the intervention on overall customer purchases. In the week immediately post-intervention, the total number of purchases, represented as $Y_{i,1}$, increase by 10% over the control group, but this effect is not statistically significant (p-value = 0.62). When we extend the observation horizon to ten weeks,[2] the average treatment effect on ten-week purchases (denoted as $Y_{i,10}$) is 0.3153 (with a p-value of 0.03), corresponding to a 30% increase (where the control group had an average of $Y_{i,10}$ of 0.99). Clearly, the company's intervention had a long-lasting impact on repeated customer purchases while the short-term impact was relatively small.

To gain deeper insights into the intervention's effects on repeated purchasing behaviors, we

[2]We use ten weeks to align with the time frame used by the focal firm when considering future purchases for newly acquired customers.

further investigate the differences in customer attrition and purchase frequency between the two treatment groups conditional on being "alive" (Figure 1.1). The leftmost figure illustrates the percentage of "alive" customers, defined as those making a purchase in a given week or afterward (up to 50 weeks after their initial purchase). The data suggests that the intervention reduced customer churn, with the treatment group consistently exhibiting higher percentages of "alive" customers than the control group. The rightmost figure in Figure 1.1 shows the weekly average number of purchases per active customer. For the first seven weeks, retained customers in the treatment group made more purchases on average than those in the control group.



Note: Customers are labeled as "alive" if they made at least one purchase in that week or later (up to week 50).

Figure 1.1: *Percentage of Alive Customers and Average Weekly Purchases After the Intervention.*

1.2.3 Designing Targeted Interventions Through CATE Estimation

The primary objective of the focal firm was to identify customer segments with the most favorable responses to the intervention with the goal of exclusively targeting these segments in future activation campaigns. With this in mind, the company aims to create a *treatment prioritization rule*, optimizing outcomes through targeted treatment assignments [12, 9, 103]. We now demonstrate how the firm can utilize the experiment to achieve this aim.

Let's begin by examining a scenario where the focal firm's goal is to maximize customer purchases within the first week post-intervention (i.e., $Y_{i,1}$). This approach aligns with prevalent coupon targeting literature [e.g., 92, 58], where the primary aim is enhancing immediate purchases post-intervention. To identify which customers should be offered additional coupons, we construct a CATE model for $Y_{i,1}$. Specifically, this model is designed to estimate the quantity $\tau_{Y_1}(\mathbf{X}_i) \equiv$

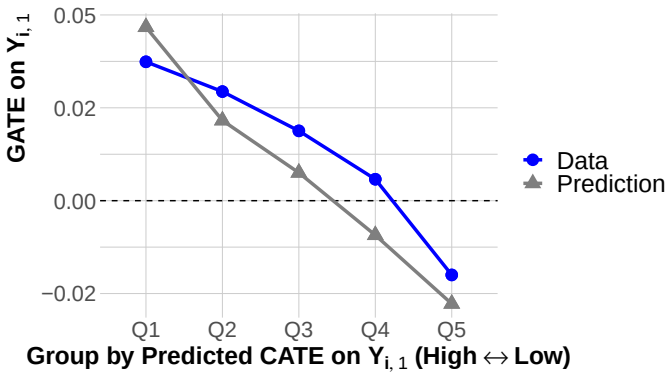
$\mathbb{E}[Y_{i,1}(1)|\mathbf{X}_i] - \mathbb{E}[Y_{i,1}(0)|\mathbf{X}_i]$, where $Y_{i,1}(W_i)$ is the potential outcome [172] of customer i's first-week purchase given the treatment condition W_i, and $\mathbf{X}_i$ includes the pre-treatment customer covariates mentioned above.

To evaluate the model's performance, we apply a bootstrap validation method similar to that used in [9]. (See Section 1.6.3 for the details.) Briefly, we first estimate a CATE model using the training set and predict CATEs for the validation customers. We then sort validation customers based on their predicted CATEs and group them into quintiles, with $\mathcal{Q}_1$ containing customers with the highest predicted CATEs (i.e., those who are predicted to increase purchases the most because of the intervention), and $\mathcal{Q}_5$ containing those with the lowest predicted CATEs. Finally, we evaluate the model's ability to identify the "right targets" by computing the *group average treatment effect* (GATE) for each quintile. We compute two measures: (i) the predicted CATEs (Prediction), and (ii) the actual outcome $Y_{i,1}$ (Data).

Figure 1.2 presents the predicted versus actual GATEs for customers in the validation dataset.[3] The closeness between predicted and actual GATEs indicates the CATE model's accuracy in estimating the true treatment effect. Specifically, the target segment recommended by the model ($\mathcal{Q}_1$) includes customers for whom the intervention was the most beneficial (with an the actual GATE of 0.037, four times the ATE), while the do-not-touch segment ($\mathcal{Q}_5$) includes customers for whom the intervention did not generate additional purchases (with an actual GATE of -0.0196). This suggests that the model can effectively rank customers according to their responsiveness to the intervention, enabling the firm to design targeted policies that maximize customer transactions within one week following the intervention.

However, the firm's primary goal is to stimulate purchases across a longer timeframe, particularly in the ten weeks following the intervention, rather than merely increasing immediate purchases. In theory, the same targeting approach should be applicable, with the *only difference* being the use of $Y_{i,10}$ as the outcome for estimating CATEs, i.e., $\tau_{Y_{10}}(\mathbf{X}_i) \equiv \mathbb{E}[Y_{i,10}(1)|\mathbf{X}_i] - \mathbb{E}[Y_{i,10}(0)|\mathbf{X}_i]$, and comparing predicted and actual GATEs. Therefore, we replicate the same analysis, this time using the total purchases made within the ten weeks following the intervention as the dependent variable.

[3]Figures 1.2 and 1.3 present the results of using the X-learner to estimate CATE. The results are consistent across various CATE models, including Causal Forest, T-learner, and S-learner, as detailed in Appendix A.4.4

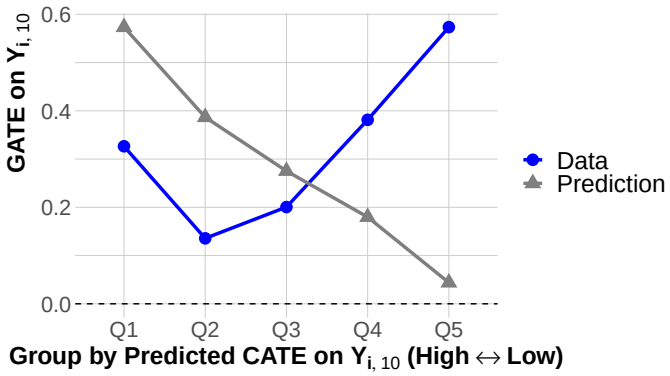


Note: Groups $\mathcal{Q}_1, \cdots, \mathcal{Q}_5$ are categorized based on the decreasing order of treatment effects predicted by the CATE model for $Y_{i,1}$. Hence, the predicted GATEs (gray line) are monotonically decreasing by definition. Actual GATEs (blue line) are computed from $Y_{i,1}$. For example, the predicted and actual GATE on $Y_{i,1}$ for $\mathcal{Q}_1$ are 0.046 and 0.038, respectively.

Figure 1.2: *Predicted and Actual GATEs by Predicted CATE Levels When the Outcome Variable is $Y_{i,1}$.*

Figure 1.3 presents the predicted versus actual GATEs on $Y_{i,10}$ for the validation customers. The U-shaped trajectory of the actual GATEs underscores the inability of the CATE model to accurately rank customers by their treatment effects on $Y_{i,10}$. For instance, if the company chooses to target customers within $\mathcal{Q}_1$ (those anticipated to show the strongest effect), the actual uplift from targeting this segment would be an increase of 0.41 purchases, contrasting the 0.58 predicted by the model. This divergence is even more apparent for customers predicted to benefit the least from the treatment (those in $\mathcal{Q}_5$). While the model predicts zero impact for this group, in reality, targeting them would result in an uplift of 0.63 purchases, notably outperforming the outcome of targeting $\mathcal{Q}_1$. Consequently, targeting strategies based on this CATE model would be ineffective.

Why does the test-to-target approach succeed for optimizing short-term outcome ($Y_{i,1}$) but struggle with long-term outcome ($Y_{i,10}$)? Is this a general phenomenon that extends beyond this particular example? If so, how can marketers design targeted interventions to optimize long-term outcomes? We address these questions in the remaining of this article. Section 1.3 examines common consumer behaviors that drive the accumulation of unexplained variations and analyzes their consequences for CATE estimation and targeting when optimizing long-term outcomes. Section 1.4 introduces a general solution that utilizes the less noisy short-term behaviors to predict the long-term treatment effect while reducing unexplained variations, along with the proposed strategy to address customer attrition. In Section 1.5, we validate our solution through



Note: Groups $\mathcal{Q}_1, \cdots, \mathcal{Q}_5$ are categorized based on the decreasing order of treatment effects predicted by the CATE model for $Y_{i,10}$. Actual GATEs (blue line) are computed from $Y_{i,10}$.

Figure 1.3: *Predicted and Actual GATEs by Predicted CATE Levels When the Outcome Variable is* $Y_{i,10}$.

simulation analyses and explore the trade-off between information gain and noise accumulation. We demonstrate the superiority of our approach in a real-world marketing campaign in Section 1.6. Finally, we conclude in Section 1.7 and suggest several research directions for future work.

1.3 The Problem: Unexplained Variations in Long-Term Repeated Purchases

The difficulty of targeting for long-term outcomes arises from two interrelated issues. First, long-term outcomes — particularly those involving repeated consumer interactions — tend to have high levels of unexplained variation due to the accumulation of unexplained customer behavior. Second, this noise accumulation problem can significantly undermine the precision of popular CATE models, resulting in suboptimal targeting approaches. Our theoretical analysis formalizes and generalizes this problem by examining two critical aspects: (i) identifying *when* and *why* the unexplained variance for long-term outcomes increases as the observation window expands, and (ii) understanding the impact of noise accumulation on the accuracy of state-of-the-art CATE models and the consequent targeting performance. (Detailed proofs of all theoretical results are available in the Appendix A.1.)

1.3.1 Characterizing Long-term Repeated Purchasing Outcomes

We examine a particular type of outcome variable: the *cumulative sum of recurring behaviors* over time. Such outcomes are commonly observed in marketing, ranging from repeated transactions in retail companies, to total engagement time for social media platforms, and customer lifetime value for SaaS businesses. For illustration throughout this paper, let's take the example of a retail company aiming to maximize total purchases over T periods. The firm plans to achieve this by targeting a marketing intervention, like distributing additional coupons to specific types of customers. Let $W_i \in \{0, 1\}$ denote the treatment assigned to customer i, and let $\mathbf{X}_i$ represent the pre-treatment covariates used for determining this assignment. The total purchases made by customer i over T periods is represented by $Y_{i,T}$. This is equivalent to the aggregate of transactions across all periods, expressed as $Y_{i,T}(W_i) \equiv \sum_{t=1}^{T} S_{i,t}(W_i)$. Here, $S_{i,t}(W_i)$ denotes the purchases by customer i in period t given the treatment W_i.

Our first objective is to investigate how the unexplained variations in $Y_{i,T}(W_i)$ — specifically the variation in $Y_{i,T}(W_i)$ that cannot be attributed to $\mathbf{X}_i$ and W_i — evolves as T increases. To achieve this, we break down the long-term outcome in the following manner:

$$Y_{i,T}(W_i) \;=\; \underbrace{\sum_{t=1}^{T} \mathbb{E}[S_{i,t}(W_i)|\mathbf{X}_i]}_{=\mathbb{E}[Y_{i,T}(W_i)|\mathbf{X}_i]} \;+\; \underbrace{\sum_{t=1}^{T} \varepsilon_{i,t}^{S}}_{\equiv \varepsilon_i^{Y_T}},$$

where $\mathbb{E}[S_{i,t}(W_i)|\mathbf{X}_i]$ is the expected purchases in period t for customer i given treatment W_i and covariates $\mathbf{X}_i$, and $\varepsilon_{i,t}^{S}$ represents the mean-zero unexplained variations of purchases in period t. Then, the variance of unexplained variations in the outcome of interest is given by:

$$\mathrm{Var}[\varepsilon_i^{Y_T}] = \mathrm{Var}\left[\sum_{t=1}^{T} \varepsilon_{i,t}^{S}\right] = \sum_{t=1}^{T} \mathrm{Var}\left[\varepsilon_{i,t}^{S}\right] + 2 \sum_{1 \leqslant t_1 < t_2 \leqslant T} \mathrm{Cov}\left[\varepsilon_{i,t_1}^{S}, \varepsilon_{i,t_2}^{S}\right].$$

This decomposition reveals that when there is non-negative serial correlation in the per-period unexplained variations (specifically, when $\mathrm{Cov}\left[\varepsilon_{i,t_1}^{S}, \varepsilon_{i,t_2}^{S}\right] \geqslant 0$ for any $t_1 < t_2 \leqslant T$), , the variance of these unexplained variations in $Y_{i,T}$ increases with the length of the observation period T. We next argue that common factors in marketing contexts, such as unobserved heterogeneity and customer attrition, often lead to positive serial correlation in these unexplained variations [e.g., 91, 75, 171]. Consequently, the existence of these factors in customer behavior typically results in

a progressive accumulation of noise over time.

Unobserved Heterogeneity.

When there are variations in customers' intrinsic preference toward the company — commonly known as *unobserved heterogeneity* or *individual fixed effects* — positive serial correlation in unexplained variations arises [e.g., 121, 87]. To illustrate that, assume that the unexplained variation in each period can be expressed as

$$\varepsilon_{i,t}^{S} = \bar{\varepsilon}_{i}^{S} + \eta_{i,t}^{S},$$

where $\bar{\varepsilon}_{i}^{S}$ represents the time-invariant individual purchase tendency (that is not captured by $\mathbf{X}_i$), and $\eta_{i,t}^{S}$ denote the independent per-period shock that is independent of $\bar{\varepsilon}_{i}^{S}$. Then, the serial correlation of $\varepsilon_{i,t}^{S}$ is positive because

$$\text{Cov}\left[\varepsilon_{i,t_1}^{S}, \varepsilon_{i,t_2}^{S}\right] = \text{Var}\left[\bar{\varepsilon}_{i}^{S}\right] + \text{Cov}\left[\bar{\varepsilon}_{i}^{S}, \eta_{i,t_1}^{S}\right] + \text{Cov}\left[\bar{\varepsilon}_{i}^{S}, \eta_{i,t_2}^{S}\right] + \text{Cov}\left[\eta_{i,t_1}^{S}, \eta_{i,t_2}^{S}\right] = \text{Var}\left[\bar{\varepsilon}_{i}^{S}\right] > 0.$$

This result emphasizes that when observable characteristics ($\mathbf{X}_i$) are insufficient to capture customer heterogeneity in the long-term outcome, the unexplained variations will exhibit positive serial correlations over time.

Customer Attrition.

Also known as customer churn, customer attrition represents the progressive loss of customers over time. This phenomenon exits in many business contexts and has been extensively analyzed in the context of recurring purchase behaviors and customer relationship management [e.g. 178, 73, 152, 11, 19]. In the following discussion, we demonstrate how customer attrition, whether observed or latent, results in a positive serial correlation in unexplained variations.

To get the intuition why customer attrition implies positive serial correlation, let's first consider the scenario where a customer churns in period t. In such a case, all future unexplained variations for this customer would be negative, given that their actual purchases reduce to zero. Conversely, if a customer remains active at time t, the unexplained variations from all earlier periods for this person would likely skew positive. This occurs because the expected per-period purchases across all customers (comprising both churned and active individuals) tend to be lower than the

realized purchases of a customer who remains active. Therefore, we can infer that the unexplained variation in per-period purchases exhibits positive serial correlations.

To formally demonstrate that customer attrition causes positive serial correlation of unexplained variations, we need to unpack the dynamics of how customer attrition affects $\varepsilon^S_{i,t}$. When a customer is still "alive" in period t, the actual purchase for that period might diverge from the expected purchase, and we denote this deviation as $\eta^S_{i,t}$. For simplicity, let's assume these deviations are independent across distinct periods (i.e., there is no unobserved heterogeneity). Conversely, for a customer who has churned in period t (or before), the unexplained variation is the negative expectation of their purchase in that period, i.e., $-\mathbb{E}[S_{i,t}(W_i)|\mathbf{X}_i]$. Consequently, the unexplained variation for period t can be expressed as:

$$\varepsilon^S_{i,t} = \mathbb{1}\left[i \text{ remains active at } t\right]\eta^S_{i,t} - \mathbb{1}\left[i \text{ has churned at } t\right]\cdot\mathbb{E}[S_{i,t}(W_i)|\mathbf{X}_i].$$

Given this formulation, the expected value of $\eta^S_{i,t}$ is positive, ensuring that $\mathbb{E}[\varepsilon^S_{i,t}] = 0$. This is consistent with the intuition above: since the expected short-term purchase, denoted as $\mathbb{E}[S_{i,t}(W_i)|\mathbf{X}_i]$, is an average taken across both alive and churned customers, it follows that the expected purchase from an "alive customer" should surpass this average.

Proceeding to the correlation dynamics, we show in Appendix A.1.1 that the covariance between the unexplained variations for periods $t_1 < t_2$ is:

$$\begin{aligned}\text{Cov}\left[\varepsilon^S_{i,t_1}, \varepsilon^S_{i,t_2}\right] = \left[1-\theta_{t_1}(W_i|\mathbf{X}_i)\right]\left\{\mathbb{E}\left[S_{i,t_1}(W_i)|\mathbf{X}_i\right] + \mathbb{E}\left[\eta^S_{i,t_1}\right]\right\}\times \\ \theta_{t_2}(W_i|\mathbf{X}_i)\left\{\mathbb{E}\left[S_{i,t_2}(W_i)|\mathbf{X}_i\right] + \mathbb{E}\left[\eta^S_{i,t_2}\right]\right\} \geqslant 0,\end{aligned} \tag{1.1}$$

where $\theta_t(W_i|\mathbf{X}_i)$ denotes the probability that the customer is still alive at t. Essentially, this covariance represents the co-movement attributed to customer attrition at time t_1, as it is the product of (i) the likelihood of customer churn at t_1 multiplied by the expected impact if the customer churns at this time (i.e., $\left[1-\theta_{t_1}(W_i|\mathbf{X}_i)\right]\left\{\mathbb{E}\left[S_{i,t_1}(W_i)|\mathbf{X}_i\right] + \mathbb{E}\left[\eta^S_{i,t_1}\right]\right\}$), and (ii) the probability of the customer remaining active at t_2 combined with the expected purchase when alive at t_2 (i.e., $\theta_{t_2}(W_i|\mathbf{X}_i)\left\{\mathbb{E}\left[S_{i,t_2}(W_i)|\mathbf{X}_i\right] + \mathbb{E}\left[\eta^S_{i,t_2}\right]\right\}$). Given that every term in (1.1) is non-negative, the resulting covariance is also non-negative.

Other Customer Behaviors.

Certainly, other behavioral factors can also contribute to positive (or negative) serial correlation. For instance, the presence of state dependence, habit persistence, and psychological switching costs can lead consumers' past consumption to positively influence their future consumption [e.g., 171, 125, 179, 59]. Under such circumstances, a notable increase in previous purchases can boost subsequent purchases, resulting in positively correlated unexplained variations. On the other hand, in settings where consumers exhibit stockpiling behavior [194], there might exist a negative correlation between unexplained variations across different periods. It will depend on whether the strength of the stockpiling behavior is stronger than that of the unexplained variations driven by unobserved heterogeneity and attrition.

To summarize, the degree of noise accumulation is determined by two main factors: (i) the presence and significance of behavioral drivers in the dataset, and (ii) the effectiveness of using observable characteristics to predict these drivers. For instance, if the observable characteristics fail to comprehensively capture customers' inherent preferences towards the company, this can lead to a significant accumulation of unexplained variations in long-term outcomes. Similarly, in situations where customer attrition is prevalent, the unexpected churn can lead to significant shocks in total purchases, resulting in increased unexplained variations in the long-term outcomes.

In our empirical application, we expect to have unobserved heterogeneity since the information from the initial purchase is unlikely to account for all the variation in individual preferences. Besides, Figure 1.1 demonstrates significant customer attrition, contributing to the accumulation of unexplained variation due to unpredictable churn. Moreover, stockpiling is unlikely in this context given the perishable nature of the goods sold in their vending machines. Taking all these factors into consideration, the focal firm is indeed facing the problem of increasing noise in their outcome of interest (i.e., long-term total purchases) as they increase the observation window. We present evidence supporting noise accumulation and positive serial correlation of unexplained variations in Section 1.6.2.

1.3.2 Implications for CATE Estimation and Targeting

The second objective of our theoretical analyses is to examine how noise accumulation in the outcome variable impacts the predictive accuracy of CATE models and their effectiveness in targeting. Within this context, we establish two key results: first, the variance in predicted CATEs escalates with increasing unexplained variations; and second, as this variance grows, the likelihood of making incorrect targeting decisions also increases.

In our analysis, we assume the firm has conducted a randomized controlled experiment such that the complete randomization, overlap, and no interference assumptions hold [117]. Additionally, we focus on a broad spectrum of CATE models, as detailed in Assumption A.1 in Appendix **??**. A defining feature of these CATE models is that the prediction for a new individual can be represented as a weighted average of residualized outcomes from the training data, where the weights are determined by the degree of similarity in covariates between individuals in the training set and the new individual and estimated using honest estimation [15]. Besides, the residualization function is constructed using cross-fitting [154]. Essentially, this class includes popular models such as Causal Forest [199], S-learners and T-learners with different outcome models [134], and R-learners with a variety of second-stage estimators [155, 126]. Appendix **??** provides a formal characterization of these models.

The following theorem formally establishes the relationship between the magnitude of unexplained variations and the variance of state-of-the-art CATE models:

Theorem 1.1 (Variance of CATE Prediction) *Assume that the CATE model (denoted as $\widehat{\tau}_{Y_T}$) belongs to the general class described above. Then, the variance of the predicted CATE for an individual with covariates $\mathbf{x}_{\text{new}}$ scales with the amount of unexplained variations in the outcome variable. Mathematically, for given the same level of observed heterogeneity in the training set (i.e., $\{\mathbf{X}_i, W_i\}_{i=1}^N$), there exists two constants C_1 and C_2 such that*

$$C_1 \text{Var}[\varepsilon_i^{Y_T}] \leqslant \text{Var}\left[\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}}) \,\middle|\, \{\mathbf{X}_i, W_i\}_{i=1}^N\right] \leqslant C_2 \text{Var}[\varepsilon_i^{Y_T}] \tag{1.2}$$

when $\text{Var}[\varepsilon_i^{Y_T}] \geqslant \sigma_0^2$ *for some* σ_0^2.

Theorem 1.1 shows that the upper and lower bounds of the variance of the predicted CATE are proportional to the variance of the unexplained variation in the outcome variable when it is

non-trivial, with both bounds increasing monotonically in $\text{Var}[\varepsilon_i^{Y_T}]$. Consequently, the unexplained variation in the outcome variable, irrespective of its impact on the consistency of CATE estimators, leads to instability in CATE models. This insight is especially crucial for practitioners focusing on optimizing long-term outcomes, as accumulating unexplained variations can significantly heighten the variance of the CATE estimator.

Next, we assess how often targeting decisions informed by the CATE model diverge from the optimal targeting strategy, which involves targeting customers with positive true CATEs. The result in Theorem 1.1 leads to the following theorem:

Corollary 1.1 (Mistargeting Probability) *Suppose the company has an unbiased CATE model $\widehat{\tau}_{Y_T}$ and targets only customers predicted to have a positive treatment effect.*[4] *Then, the probability of the model deviating from the optimal policy, $\mathbb{P}\left[\tau_{Y_T}(\mathbf{x}_{\text{new}}) \cdot \widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}}) < 0\right]$, increases with the variance of the predicted CATE, i.e., $\text{Var}\left[\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}})\right]$.*

Corollary 1.1 reveals a fundamental challenge in targeting for long-term outcomes: instability in CATE predictions caused by unexplained variations in the outcome variable can result in erroneous targeting decisions. Therefore, regardless of the application domain, companies will invariably encounter challenges in accurate targeting when faced with substantial unexplained variations in the outcome variable.

A possible solution to the high variance problem in optimizing long-term outcomes is the use of CATE estimators specifically tailored to reduce prediction variance. This can be achieved by constraining the complexity of the CATE model, such as applying lasso regularization during the estimation process of R-learners [155]. However, while these estimators are effective at reducing variance, they often introduce substantial underfitting bias. Our empirical analyses indicate that in cases with limited sample sizes, this bias can significantly impair targeting accuracy (refer to Table 1.1 for empirical results). Therefore, the effective use of regularization necessitates large sample sizes.

Finally, there are alternative approaches for developing targeting policies. These methods, known as policy learning, directly leverage experimental data to learn the optimal allocation

[4]This proposition also holds in the case where the firm aims to target customers with treatment effects larger than a certain threshold. See Appendix A.1 for details.

of treatments. They typically derive a proxy variable for the true CATE, such as the inverse probability weighting [133] or doubly robust scores [17], and then estimate a policy (often using machine learning models) to determine which customers the firm should target. Since these methods depend on a proxy variable derived from experimental data, they are conceptually similar to the targeting approach examined in this section. Therefore, they encounter a similar challenge: substantial unexplained variations in the outcome variable can increase the variance of the CATE proxy, resulting in ineffective targeting policies. We provide additional empirical evidence in Section 1.6.4.

1.3.3 Summary

Our theoretical analysis reveals that for outcomes based on repeated customer behaviors, factors like unobserved heterogeneity and customer churn are key contributors to unexplained per-period variations. Over time, these variations accumulate, resulting in increased variance in outcomes as the observation period lengthens. As a result, the accuracy of conventional CATE models is compromised due to this noise accumulation, leading to less effective targeting decisions. These insights highlight a critical challenge in estimating CATEs and developing effective targeting strategies for long-term business outcomes.

1.4 The Solution: Surrogate Index with Separate Imputation

So far, we have identified that the primary issue with ineffective long-term targeting arises from the accumulation of unexplained variations in the outcome of interest. Therefore, a potential solution to improve targeting effectiveness is to reduce the unexplained variations in long-term outcomes that are not attributable to the intervention. To achieve this, we suggest using a *noise-reduced proxy* in place of the actual outcome variable when estimating the CATE model. This proxy aims to capture the long-term treatment effect heterogeneity while excluding the variations unrelated to the intervention.

1.4.1 Solution Overview

We use the *surrogate index* [e.g., 16, 212] to construct the noise-reduced proxy. This index represents the expected long-term outcome, derived from a set of *observed short-term outcomes* — such as immediate post-intervention purchases in our case — and pre-treatment covariates. To construct a surrogate index, companies can use historical data to build a model that identifies the relationship between short-term behaviors (along with customer covariates) and the actual long-term outcome. After establishing such a model, companies can collect short-term data, use it to predict long-term outcomes, and then assess the impact of their intervention based on these predicted long-term results. This approach offers three major advantages:

1. *Noise Reduction*: By design, the surrogate index incorporates only unexplained variations from the short-term outcomes in the surrogate model. Therefore, the predicted long-term outcome (or proxy) excludes the unexplained variations of behaviors occurring after these short-term outcomes were collected. As a result, the surrogate index includes *fewer unexplained variations* (i.e., those unrelated to the intervention) compared to the actual long-term outcome.

2. *Long-term Orientation*: The surrogate index effectively represents the long-term effect of the intervention. Many marketing interventions result in long-term shifts in consumer behaviors by initiating immediate behavioral modifications. In such scenarios, short-term outcomes frequently account for a significant portion, if not all, of the long-term treatment effect. For example, the coupon promotion in our empirical application can reduce customer churn in the initial week following the intervention, as illustrated in Figure 1.1, and this enhanced retention eventually results in increased aggregate long-term purchases due to the extended lifetime of customers who received more promotions. The short-term purchase data in this scenario reflects the enhanced retention, enabling the surrogate index to use this information to forecast the sustained, long-term retention improvements.

3. *Acceleration of Decision Time*: For constructing a surrogate index, companies require: (i) a model that links short-term signals with long-term outcomes, which is estimated using historical data; (ii) pre-intervention covariates that are available before implementing

the intervention; and (iii) short-term outcomes that are observed immediately following the intervention. As a result, companies can develop targeting strategies for optimizing long-term outcomes soon after the intervention (i.e., as soon as the short-term signals are observed), thus bypassing the delay associated with waiting for observing the actual long-term outcomes.

The first benefit of the surrogate index method highlights its potential to improve long-term targeting effectiveness. By reducing unexplained variations in the outcome variable, it enhances CATE estimation precision (as per Theorem 1.1), leading to more accurate targeting decisions (Corollary 1.1). The second benefit confirms the validity of using the surrogate index for estimating long-term treatment effects: although such targeting policies mainly depend on short-term behaviors, they can be effective if these behaviors explain a significant proportion of the long-term treatment effect. The third benefit, although not essential in our case since the focal company can postpone the implementation of new targeting rules, might be crucial in situations where rapid decision-making can lead to substantial savings in organizational costs.

We next provide the formal definition of a surrogate index, specify the conditions required for it to capture long-term treatment effects (benefit #2), and illustrate its variance reduction property (benefit #1).

1.4.2 Identification and Variance Reduction Using Surrogate Index

Assume that the company has access to two datasets: the experimental data with the intervention (denoted as $\mathcal{E}$), and the historical data without the intervention (denoted as $\mathcal{H}$). The *surrogate index* is defined as follows:

Definition 1.1 (Surrogate Index) *The surrogate index is the expected long-term outcome ($Y_{i,T}$) of customers in $\mathcal{H}$, conditioned on their short-term behaviors ($\mathbf{S}_{i,T_0} = \{S_{i,1}, \cdots, S_{i,T_0}\}$ for some $T_0 < T$) and pre-treatment covariates ($\mathbf{X}_i$). Mathematically, it can be represented as $\widetilde{Y}_T(\mathbf{S}_{T_0}, \mathbf{X}_i) \equiv \mathbb{E}_{\mathcal{H}}\left[Y_{i,T} | \mathbf{S}_{T_0}, \mathbf{X}_i\right]$.*

To ensure identification of CATEs using the short-term signals, the following assumptions are made [13]:

Assumption 1.1 (Identification Assumptions for Long-term CATEs)

1. (Surrogacy) *The short-term outcomes can fully mediate the treatment effect of W_i on $Y_{i,T}$; that is,* $W_i \perp\!\!\!\perp Y_{i,T} \mid \mathbf{S}_{i,T_0}, \mathbf{X}_i \ , \ \forall i \in \mathcal{E}$.

2. (Comparability) *The experimental and historical data are comparable in distribution; that is,* $Y_{i,T} \mid \mathbf{S}_{i,T_0}, \mathbf{X}_i \ , \ i \in \mathcal{E} \ \overset{d}{\sim} \ Y_{i,T} \mid \mathbf{S}_{i,T_0}, \mathbf{X}_i \ , \ i \in \mathcal{H}$.

The surrogacy assumption states that the variations in short-term behaviors induced by the intervention can fully reflect its causal impact on the long-term outcome. Therefore, if this assumption holds and we can accurately estimate the influence of $\mathbf{S}_{i,T_0}$ on $Y_{i,T}$, then the long-term treatment effect can be deduced simply by observing the short-term outcomes of the experimental units. The comparability assumption ensures that the impact of $\mathbf{S}_{i,T_0}$ on $Y_{i,T}$ is the same for both the experimental data and the historical data, which implies that one can use the historical data to infer the impact of $\mathbf{S}_{i,T_0}$ on $Y_{i,T}$ and then extrapolate this relationship to the experimental data.

Altogether, when both of these assumptions hold true, the surrogate index representation becomes a reliable tool for estimating the CATE on the long-term outcome. Furthermore, since the surrogate index is based on short-term outcomes, it contains less unexplained variation compared to the actual long-term outcome. Formally, we state the theorem as follows:

Theorem 1.2 (Identification and Variance Reduction Using Surrogate Index) *Suppose that Assumption 1.1 holds.*

1. *The CATE of the intervention on the long-term outcome is equal to the CATE on the surrogate index, that is,* $\tau_{Y_T}(\mathbf{X}_i) = \widetilde{Y}_T(\mathbf{S}_{i,T_0}(1), \mathbf{X}_i) - \widetilde{Y}_T(\mathbf{S}_{i,T_0}(0), \mathbf{X}_i)$, *where* $\mathbf{S}_{i,T_0}(W_i)$ *denotes the potential outcome of the short-term outcomes.*

2. *The variance of the surrogate index is smaller than the variance of the actual long-term outcome, that is,* $\text{Var}\left[\widetilde{Y}_T(\mathbf{S}_{i,T_0}(W_i), \mathbf{X}_i)\right] < \text{Var}[Y_{i,T}(W_i)|\mathbf{X}_i]$.

Theorem 1.2.1 suggests that the surrogate index offers an unbiased estimation of the CATE for the long-term outcome, assuming we can correctly estimate $\mathbb{E}_{\mathcal{H}}\left[Y_{i,T}|\mathbf{S}_{T_0}, \mathbf{X}_i\right]$. Meanwhile, Theorem 1.2.2 demonstrates that the surrogate index has lower variance compared to the actual long-term outcome. Combining these insights with those from Theorem 1.1 and Corollary 1.1, it becomes evident that employing the surrogate index method for CATE estimation can (i) reduce

variance in CATE predictions, and (ii) decrease the likelihood of mistargeting when the objective is optimizing the long-term outcome.

Importantly, the realization of these benefits depends primarily on the validity of Assumption 1.1, as discussed in Section 1.4.4. It also relies on the firm's capability to accurately estimate $\mathbb{E}_{\mathcal{H}}\left[Y_{i,T}|\mathbf{S}_{T_0}, \mathbf{X}_i\right]$ using available historical data, which we will discuss in the following section.

1.4.3 Separate Imputation Approach

Conventionally, researchers have constructed the surrogate index by employing a regression model that associates long-term outcomes with short-term results and pre-intervention covariates [e.g., 16, 212]. However, in many marketing scenarios, this method may encounter a challenge of high variance, particularly in the presence of *customer attrition*, as unexplained customer churn hinders the model's ability to distinguish between variations that can and cannot be explained by the short-term outcomes. In this section, we delve into the reasons for this shortcoming and introduce an novel solution that firms can seamlessly adopt when constructing the surrogate index.

Challenge: Inseparable Unexplained Variations Arising from Attrition.

When customer attrition exists, the total purchases of customer i over T periods ($Y_{i,T}$) are determined by two factors: the customer's active lifetime up to period T (denoted as $\mathcal{T}_i^T$), and their expected purchase per period while active (denoted as Λ_i^T). We can further break down these factors into variations that can be explained by $\mathbf{S}_{i,T_0}$ and $\mathbf{X}_i$, and unobserved variations arising from individual preferences towards the firm not captured in the data (e.g., unobserved heterogeneity, random shocks, etc.). Hence, the aggregate purchase counts for customer i can be expressed as:

$$\begin{aligned} Y_{i,T} &= \mathcal{T}_i^T \times \Lambda_i^T = \left\{\mathbb{E}\left[\mathcal{T}_i^T|\mathbf{S}_{i,T_0}, \mathbf{X}_i\right] + \varepsilon_i^{\mathcal{T}}\right\} \times \left\{\mathbb{E}\left[\Lambda_i^T|\mathbf{S}_{i,T_0}, \mathbf{X}_i\right] + \varepsilon_i^{\Lambda}\right\} \\ &= \mathbb{E}\left[\mathcal{T}_i^T|\mathbf{S}_{i,T_0}, \mathbf{X}_i\right]\mathbb{E}\left[\Lambda_i^T|\mathbf{S}_{i,T_0}, \mathbf{X}_i\right] + \underbrace{\mathbb{E}\left[\Lambda_i^T|\mathbf{S}_{i,T_0}, \mathbf{X}_i\right]\varepsilon_i^{\mathcal{T}} + \mathbb{E}\left[\Lambda_i^T|\mathbf{S}_{i,T_0}, \mathbf{X}_i\right]\varepsilon_i^{\Lambda}}_{\equiv \xi_i(\mathbf{S}_{i,T_0}, \mathbf{X}_i),\ \text{additional variations}} + \varepsilon_i^{\mathcal{T}}\varepsilon_i^{\Lambda}, \end{aligned} \tag{1.3}$$

where $\varepsilon_i^{\mathcal{T}}$ and ε_i^{Λ} denote the unexplained variations in lifetime and purchase intensity, respectively.

The above formulation reveals a key challenge in estimating the relationship between short-term and long-term outcomes: there is an additional term, $\xi_i(\mathbf{S}_{i,T_0}, \mathbf{X}_i)$, which connects explained variations in customer lifetime (attributable to short-term outcomes and observed covariates) with unexplained variations in per-period purchase intensity, and vice versa. This implies that directly modeling $Y_{i,T}$ using $\mathbf{S}_{i,T_0}$ (and $\mathbf{X}_i$) becomes problematic, as the model may not effectively differentiate whether variations in $Y_{i,T}$ are driven by $\mathbf{S}_{i,T_0}$ (and $\mathbf{X}_i$) or by $\varepsilon_i^{\mathcal{T}}$ and ε_i^{Λ}. This inseparability results in *high-variance* in the surrogate model, as it struggles to separate the explainable and unexplainable variations present in the historical data. (Further discussion and an empirical example of this phenomenon are available in the Appendix **??**.)

Previous research has addressed the inseparability problem by introducing additional assumptions [31, 169, 44, 105, 116]. For example, a common method to tackle the multiplicative noise structure is to assume that the unexplained part in the outcome variable (e.g., $\xi_i(\mathbf{S}_{i,T_0}, \mathbf{X}_i)$ and $\varepsilon_i^{\mathcal{T}} \varepsilon_i^{\Lambda}$ in (1.3)) is independent of the target variables ($\mathbf{S}_{i,T_0}$ in our case) after controlling for observed covariates [e.g. 105, 106, 187]. However, this assumption does not hold in our scenario since $\xi_i(\mathbf{S}_{i,T_0}, \mathbf{X}_i)$ is clearly associated with $\mathbf{S}_{i,T_0}$ even after controlling for $\mathbf{X}_i$. Another solution proposed in the literature is to employ instrumental variables that correlate with $\mathbf{S}_{i,T_0}$ but remain unassociated with the unexplained variations [42]. Yet, this approach is rarely feasible in practice since it necessitates historical data that includes exogenous shocks capable of serving as instrumental variables when modeling the relationship between short-term and long-term outcomes.

Consequently, we propose a new imputation technique for creating the surrogate index, specifically designed to tackle the inseparability problem stemming from customer attrition.

Solution: Separate Models for Customer Attrition and Purchase Intensity.

Upon examining (1.3), it becomes clear that the issue of inseparable unexplained variations can be addressed by separately estimating two relationships: first, between customer lifetime and short-term outcomes (i.e., $\mathbb{E}\left[\mathcal{T}_i^T | \mathbf{S}_{i,T_0}, \mathbf{X}_i\right]$), and another between purchase intensity and short-term outcomes (i.e., $\mathbb{E}\left[\Lambda_i^T | \mathbf{S}_{i,T_0}, \mathbf{X}_i\right]$), and combine the predictions of these two models afterwards. We refer to this as the *separate imputation* approach as we determine the long-term outcome by combining predictions from two distinct surrogate models rather than making a direct prediction.

To be more specific, the separate imputation capitalizes on the measurable nature of both customer lifetime an purchase intensity. The first model predicts customer lifetime using $\mathbf{S}_{i,T_0}$ and $\mathbf{X}_i$ (denoted by $\widehat{\mathcal{T}}_T\left(\mathbf{X}_i|\mathbf{S}_{i,T_0}\right)$), while the second one estimates the purchase intensity using $\mathbf{S}_{i,T_0}$ and $\mathbf{X}_i$ (denoted by $\widehat{\Lambda}_T\left(\mathbf{X}_i|\mathbf{S}_{i,T_0}\right)$). After constructing the two surrogate models using the historical data, we predict the lifetime and purchase intensity for customers in the experimental data using their observed short-term outcomes ($\mathbf{S}_{i,T_0}(W_i)$) and pre-treatment characteristics ($\mathbf{X}_i$). We then combine these predicted values by multiplication to create the surrogate index that will be used for CATE estimation and policy learning, i.e.,

$$\widehat{\widetilde{Y}}_T^{\text{Sep}}\left(\mathbf{X}_i|\mathbf{S}_{i,T_0}(W_i)\right) = \widehat{\mathcal{T}}_T\left(\mathbf{X}_i|\mathbf{S}_{i,T_0}(W_i)\right) \times \widehat{\Lambda}_T\left(\mathbf{X}_i|\mathbf{S}_{i,T_0}(W_i)\right),\ i \in \mathcal{E}.$$

By design, the estimation of $\widehat{\widetilde{Y}}_T^{\text{Sep}}\left(\mathbf{X}_i|\mathbf{S}_{i,T_0}(W_i)\right)$ is unaffected by $\xi_i(\mathbf{S}_{i,T_0}, \mathbf{X}_i)$, circumventing the high-variance issue stemming from the multiplicative noise pattern seen in (1.3).

Implementing Separate Imputation in Practice.

Our approach can be applied to various business contexts as long as the firm has access to historical data $\mathcal{H}$ that can be used to calibrate the two surrogate models. In contractual settings, where customer churn is observable, creating two separate models is straightforward, as both customer lifetime and average purchases per alive period can be directly calculated from the available data. In non-contractual settings, such as the one in our empirical application, customer churn is not immediately apparent, which means that the actual customer lifetime cannot be directly obtained from the data. In such contexts, we utilize established metrics like *recency* and *frequency* (derived from observed customer behaviors) to approximate customer lifetime and purchase intensity. In our application, we employ the *last observed purchase time* up to T as a proxy for $\mathcal{T}_i^T$, and the *average number of purchases per period* up to $\mathcal{T}_{i,T}$ as a surrogate for the average purchase intensity Λ_i^T.

While $\mathcal{T}_i^T$ and Λ_i^T may not flawlessly represent the true customer lifetime or purchase intensity (as customers might churn after the last period of purchase), these metrics are useful proxies because: (i) together, they provide sufficient information to infer customer lifetime and purchase intensity [73, 74]; and (ii) $\mathcal{T}_i^T$ is largely reflective of customer lifetime (for example, customers who

churn early typically show low recency), while Λ_i^T correlates strongly with purchase intensity (such as high average purchases per period indicating consistent frequent buyers). Therefore, employing these proxies for estimating surrogate index is a justified and practical approach to address the challenge of inseparability in non-contractual settings.

To conclude, using $\widehat{\approx}^{\text{Sep}}_{Y_T}(\mathbf{X}_i|\mathbf{S}_{i,T_0}(W_i))$ as a substitute for the actual outcome in CATE estimation presents significant benefits for addressing the challenges of long-term targeting. Firstly, this approach results in less noise compared to using the actual long-term outcome, as it omits unpredictable variations that occur between T_0+1 and T. Secondly, it enables valid inferences about the long-term treatment effect under Assumption 1.1. Lastly, employing separate models for churn and purchase helps to more accurately assess the influence of short-term behavioral changes on long-term purchasing patterns. As illustrated in Sections 1.5 and 1.6, this separate imputation method outperforms other strategies commonly used by firms, including existing imputation techniques.

1.4.4 Potential Limitations and Implementation Considerations

Although the surrogate index method presents significant advantages, it is important to acknowledge its limitations and the potential challenges encountered during implementation. Understanding these factors is essential for assessing the generalizability of the proposed approach and its applicability in different situations.

Existence of Less Noisy Surrogate Variables.

Similar to [13], we use immediate outcomes per period as the surrogate variables in our empirical study. These variables can reflect potential long-term impact in two ways. First, being integral parts of $Y_{i,T}$, they inherently capture parts of the long-term treatment effect. Second, changes in short-term behaviors can be indicative of shifts in long-term purchasing trends. For example, an intervention that enhances customer retention, thereby increasing long-term purchases, would likely result in a greater proportion of customers in the treatment group making non-zero short-term purchases due to reduced churn. Similarly, if the intervention encourages sustainable purchasing habits, an increase in short-term purchase frequency among treated

customers would be observed, signaling the habit formation influenced by the intervention.

Surrogate variables of this kind are particularly beneficial in contexts where the outcome involves recurring activities such as repeat purchases or engagement, commonly seen in industries like retail, subscription media, food service, and transportation. However, for sectors characterized by long purchase cycles, like the automotive industry, finding valid surrogate variables can be challenging. In these cases, our method may be less applicable, as it can be difficult to detect short-term indicators. Furthermore, when short-term outcomes are also noisy (e.g., when the data is obscured for privacy protection), the benefits of a surrogate index may not be significant. In such situations, the surrogate index method is not applicable given the lack of effective surrogate variables.

Choice of Surrogates.

The surrogacy assumption requires that short-term outcomes *fully mediate* the long-term treatment effect. When this condition is not met, the CATEs identified using the surrogate index may diverge from the actual CATEs, leading to potential mistargeting errors [212]. One way to mitigate the risk of surrogacy violation is to include a large number of surrogates [13]. For example, when addressing long-term outcomes such as repeated transactions, incorporating more periods into the surrogate index reduces the chances of violating this assumption. However, adding more periods also increases unexplained variations, which may decrease the effectiveness of targeting policies. This issue is formally characterized in the following proposition:

Corollary 1.2 (Noise Accumulation of Surrogate Indices) *Suppose that Assumption 1.1 holds. When we build two surrogate indices based on different periods of short-term outcomes, where $T_0 > T_0'$, it follows that:*

$$\text{Var}\left[\widetilde{Y}_T(\mathbf{S}_{i,T_0}(W_i), \mathbf{X}_i)\right] > \text{Var}\left[\widetilde{Y}_T(\mathbf{S}_{i,T_0'}(W_i), \mathbf{X}_i)\right].$$

Corollary 1.2 and the surrogacy assumption highlight the trade-off between information gain and noise accumulation for optimal targeting performance. Specifically, including more periods of short-term outcomes improves the validity of the surrogacy assumption. However, it also introduces unexplained variations which in turn increase the mistargeting probability. Determining the optimal number of periods for estimating surrogate models requires empirical investigation,

which is challenging due to the difficulty of directly testing the surrogacy assumption. In this paper, we address this question by comparing holdout targeting performance across surrogate indices constructed using varying numbers of periods. Our simulations and real-world data analysis indicate that surrogate models with fewer periods may significantly improve targeting effectiveness, even though they might risk violating the surrogacy assumption.

Violation of Comparability Assumption.

In this work, we construct surrogate indices using historical data. This method is based on the assumption that the relationships between short-term and long-term outcomes remain consistent across both historical and experimental datasets. This assumption is expected to be valid in our empirical analysis, as our experiment includes *all* newly acquired customers, who share characteristics with those in the historical dataset, and the company's product offerings and acquisition strategies remained unchanged. However, if the experiment targets a specific subset of customers (like those acquired through a particular channel), additional adjustments may be necessary to ensure the surrogate models, developed from historical data, remain generalizable [e.g., 149, 173].

1.5 Empirical Performance: Simulation Evidence

1.5.1 Simulation Setting

We first validate our solution using synthetic data. Our simulations consider a company implementing a marketing intervention with the aim to maximize total purchases ($Y_{i,T}$) over a ten-week period ($T = 10$) post-intervention. We generate experimental data ($\mathcal{E}$) with accumulated unexplained variations over time. Specifically, we assume that the intervention has a direct and heterogeneous impact on churn probability and purchase intensity during the first three weeks, and no further impact after the fourth week. For simplicity, we assume that the unexplained variations in purchase intensity (while alive) are independent across time. Therefore, the noise accumulation is mainly driven by unexplained customer attrition. More details regarding the simulation setting and the noise accumulation behaviors can be found in Appendices A.3.1

and A.3.2.

We also generate historical data ($\mathcal{H}$) for 5,000 customers using the data generating process for the control group, which reflects the company not implementing the intervention in the past. This data will be used to construct the surrogate models. For the main analysis, we select $T_0 = 3$ periods for constructing the surrogate index, ensuring compliance with the surrogacy assumption (Assumption 1.1). Additionally, we vary the number of periods in constructing the surrogacy index to explore the balance between gaining information and accumulating noise. The model specifications for creating surrogate indices are detailed in the Appendix A.3.5.

1.5.2 Comparison Methods

Alternative Imputation Methods.

We compare our separate imputation approach to various imputation methods for surrogate index construction. The first technique we examine is the single imputation approach proposed by [13] and [212]. In this method, the outcome variable ($Y_{i,T}$) is regressed directly onto the short-term signals ($S_{i,1}, \cdots, S_{i,T_0}$) as well as the pre-intervention covariates ($\mathbf{X}_i$). However, this imputation approach does not account for the inseparable nature of unexplained variations stemming from customer attrition. As a result, this model offers a less accurate estimation for the impact of short-term outcomes on $Y_{i,T}$ compared to the separate imputation technique.

We also examine the BG/NBD model with covariates [71] as an alternative imputation approach. This model has not been traditionally suggested as an imputation technique in the surrogate index literature, but it could be a reasonable candidate in our context because it has been shown to be effective in capturing unobserved heterogeneity in customer attrition and purchase intensity. Note that the accuracy of the BG/NBD model, along with similar variants like the Pareto/NBD model, heavily depends on the distributional assumptions and the assumed functional forms of the relationship between customer covariates. Consequently, in situations where the relationship between observed covariates and the heterogeneity of treatment effects is intricate, we expect the BG/NBD model to underperform when contrasted with more flexible frameworks, such as non-linear regressions or machine learning models.

Alternative Variance Reduction Methods.

In addition to using alternative imputation methods to obtain the surrogate index, we investigate other techniques to reduce the variance in CATE estimation. One such technique is to explicitly regularize the CATE function during the estimation process. Here, we use R-learner with lasso regularization when estimating CATE function [155] to study the effectiveness of regularization as a potential solution. Although regularization can reduce the variance of CATE models, it can also introduce significant underfitting bias [97] — the penalty term from regularization may cause the model to overlook crucial data patterns, which can be particularly problematic when dealing with a small training sample with large unexplained variations.

Another obvious alternative to reduce variance is to increase the sample size of the experiment data. While this is straightforward to evaluate when using synthetic data, this is not a feasible or optimal solution for enhancing targeting policies in practice. First, the number of customers who qualify for the intervention is often limited, which limits the sample size available to most firms.[5] Second, even when companies can increase the experimental sample size, the rate at which the variance of CATE decreases with respect to sample size can be considerably slow when the noise level is high. Nonetheless, we incorporate this alternative approach in our simulation analysis to evaluate the potential advantages of scaling the experiment, as opposed to employing different dependent variables for CATE estimation.

Baseline Approaches.

Finally, we examine two baseline methods commonly used in practice: the *default* approach and the *myopic* approach. The default approach targets customers based on their actual long-term outcome, $Y_{i,T}$. This approach is the simplest and most common, but it is expected to be ineffective due to the substantial unexplained variations in $Y_{i,T}$. The myopic approach targets customers based on their short-term performance, $Y_{i,T_0} = \sum_{t=1}^{T_0} S_{i,t}$ (i.e., based on their behavior only a few periods right after the intervention). This approach can avoid noise accumulation (as there is less unexplained variations in behavior up to T_0), but it may not yield optimal performance because it disregards the disparities between short-term and long-term treatment effects.

[5] [185] propose an approach to calculate the sample size required to train and certify targeting policies.

1.5.3 Evaluation Procedure

We assess targeting performance through 200 bootstrap replications and report the mean and standard deviation of key metrics. In each replication, we first create a training and a validation set. For each of the approaches considered, we use the training set to construct a CATE model $\widehat{\tau}_{\ddot{Y}}(\mathbf{X}_i)$ using $\ddot{Y}$ as the dependent variable (e.g., $\ddot{Y} = Y_{10}$ for the default approach, $\ddot{Y} = \widehat{\widehat{Y}}_T^{\text{Sep}}$ for the proposed approach, etc.). We then calculate the area under the targeting operating characteristic curve (AUTOC) [210] using the actual long-term outcome ($Y_{i,T}$) on the validation set.

Specifically, AUTOC is constructed as follows. Given the predicted CATEs $\widehat{\tau}_{\ddot{Y}}(\mathbf{X}_i)$, the *targeting operator characteristic* (TOC) for $Y_{i,T}$ is defined as

$$\text{TOC}(\phi; \widehat{\tau}_{\ddot{Y}}) = \mathbb{E}\left[Y_{i,T}(1) - Y_{i,T}(0) | F_{\widehat{\tau}_{\ddot{Y}}}(\widehat{\tau}_{\ddot{Y}}(\mathbf{X}_i)) \geqslant 1 - \phi\right] - \mathbb{E}\left[Y_{i,T}(1) - Y_{i,T}(0)\right],$$

where $F_{\widehat{\tau}_{\ddot{Y}}}$ is the cumulative distribution function of the predicted CATEs. The TOC measures the incremental gains from targeting the top $\phi \times 100\%$ customers, as the difference in ATE between customers in the top $\phi \times 100\%$ CATE group and all customers. Then, the AUTOC is defined as

$$\text{AUTOC}(\widehat{\tau}) = \int_0^1 \text{TOC}(\phi; \widehat{\tau}) d\phi.$$

Note that a model $\widehat{\tau}_{\ddot{Y}}$ is better than another model $\widehat{\tau}_{\ddot{Y}'}$ in identifying customers in the top $\phi \times 100\%$ CATE group if $\text{TOC}(\phi; \widehat{\tau}_{\ddot{Y}}) > \text{TOC}(\phi; \widehat{\tau}_{\ddot{Y}'})$. The AUTOC is a useful metric for evaluating the effectiveness of a CATE model because it quantifies how well the model ranks units based on their treatment effect, with a higher AUTOC indicating a more effective targeting (or treatment prioritization) rule.

1.5.4 Results

Table 1.1 shows the AUTOC values of different approaches. Each row corresponds to CATE models for a specific outcome variable, and the columns indicate the two methods used for CATE estimation and the two training sample sizes ($N = 1,000$ and $N = 50,000$). In Appendix A.3.6, we also provide the results from other CATE models (including S-learner and T-learner) to corroborate that our findings are not driven by a particular CATE model.

In the scenario of a small sample size ($N = 1,000$), there are several important findings. Firstly,

all the methods that employ short-term proxies as the dependent variable for CATE estimation have superior AUTOC performance (both higher mean values and smaller standard deviations) than the default approach. This implies that relying on short-term signals rather than the actual long-term outcome can significantly improve the effectiveness of targeting. Secondly, the separate imputation method consistently achieves the highest performance (highest AUTOC means with smallest standard deviations), regardless of models being used to estimate CATEs. In contrast, the single imputation performs worse than other short-term approaches. This finding highlights the importance of separating churn and purchase when creating a surrogate index. Thirdly, although the BG/NBD technique fares better than using the actual outcome, it falls short of the performance achieved by the separate imputation method. This can be attributed to the BG/NBD approach being less efficient when the relationship between observed characteristics and key parameters of interest is complex, as it is the case in our simulation.

Table 1.1: *Comparison of AUTOC Values for Different Outcomes and CATE Models.*

Outcome $\ddot{Y}$	$N = 1,000$		$N = 50,000$	
	Without Regularization	**With Regularization**	**Without Regularization**	**With Regularization**
Separate Imputation	**0.88 (0.04)**	**0.88 (0.13)**	**0.92 (0.02)**	**0.93 (0.02)**
Single Imputation	0.83 (0.09)	0.63 (0.40)	0.91 (0.02)	0.91 (0.02)
BG/NBD Imputation	0.84 (0.07)	0.78 (0.30)	0.91 (0.02)	0.93 (0.02)
Myopic ($Y_{i,3}$)	0.85 (0.06)	0.81 (0.28)	0.91 (0.02)	0.93 (0.02)
Default ($Y_{i,10}$)	0.55 (0.30)	0.31 (0.45)	0.73 (0.04)	0.92 (0.02)

Note: Higher AUTOC reflects better prioritization rule. We average the results over 200 replications and show in parentheses the standard deviation. We use Causal Forest as the CATE model without regularization and R-learner with lasso regularization as the model with regularization. The performance of different CATE models is provided in Appendix A.3.6.

Next, we highlight the efficiency of the proposed solution with respect to sample size. In situations where separate imputation is applied to small experimental data ($N = 1,000$), its performance (AUTOC $= 0.88$ for both methods) nearly matches the performance observed with fifty-time larger datasets (AUTOC $= 0.92$ without regularization and AUTOC $= 0.93$ with regularization). Notably, the difference in AUTOCs between the small and large datasets is lower for the separate imputation method compared to other strategies, which highlights the sample size efficiency of our proposed solution. Furthermore, when employing a non-regularized CATE model, using short-term proxies for targeting within a small experimental dataset yields better

performance with AUTOC scores between 0.83 and 0.88, compared to the standard approach trained on a significantly larger dataset, which achieves an AUTOC of 0.73. Taken together, these results make a strong case for the separate imputation technique, highlighting its ability to achieve near-optimal performance even with significantly fewer data points compared to larger datasets.

Finally, it is noteworthy that the performance of the R-lasso remains consistent across all methods when working with a large experimental dataset ($N = 50,000$). This result suggests that when there is an abundance of samples, regularization can effectively address the noise accumulation challenge. However, in smaller sample settings, applying regularization to CATE models for the default approach can lead to a significant drop in targeting performance, with the AUTOC value plummeting to 0.31. Such an outcome is driven by R-lasso's inclination to underestimate treatment effect heterogeneity, particularly in contexts with limited data.[6] Hence, firms should not depend exclusively on regularization for mitigating noise accumulation — its effectiveness may realize only when dealing with large-scale experiments.

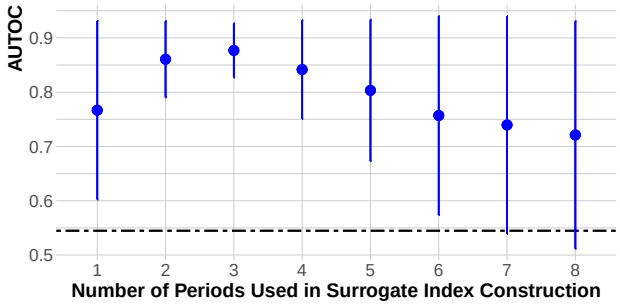
1.5.5 The Trade-off between Information Gain and Noise Accumulation

In the previous section, we emphasize cases where we incorporate only the minimal set of short-term outcomes ($T_0 = 3$) to meet the surrogacy assumption (Assumption 1.1.1). However, verifying the surrogacy assumption in real-world applications is challenging, and companies must balance between information capture and noise accumulation when determining the number of periods to include in the surrogate index. While incorporating more periods into the surrogate model could help fulfill the surrogacy assumption, it might also increase unexplained variations, potentially diminishing the targeting effectiveness (as discussed in Section 1.4.4).

We explore this trade-off with our simulated data, simulating a real-world situation where firms determine the number of surrogates to include in their surrogate model. Specifically, we construct eight distinct surrogate models, each utilizing a different number of short-term outcomes (ranging from $T_0 = 1$ to $T_0 = 8$) using the separate imputation approach. Following this, we generated 200 bootstrap replications and report the average and standard deviations of the

[6]To provide additional context, in approximately 60% of the bootstrap replications, the default method of R-lasso generated the same CATE prediction (which was equal to the ATE) for all consumers. This suggests a notable underestimation of the heterogeneity in treatment effects.

AUTOC for each model. Figure 1.4 presents the results corresponding to the number of periods ranging from $T_0 = 1$ to $T_0 = 8$.



Note: Each point reports the average over 200 simulation replications together with the one standard deviation interval. The dashline represents the mean AUTOC of the default approach. We used 1,000 customers in the training set. We present here the results of using Causal Forest, but our findings are robust across different CATE models. See Appendix A.3.8 for the results of different CATE models.

Figure 1.4: *Trade-off between Information Gain and Noise Accumulation: An Analysis of Causal Forest AUTOCs with Surrogate Index Constructed Using Different Periods.*

The inverted U-shaped relationship in Figure 1.4 reflects the trade-off between information gain and noise accumulation. As we increase T_0 from one to three periods, the AUTOC improves because the intervention has a direct impact until the third week. However, the AUTOC starts to decline once we include behaviors beyond the third period (after satisfying surrogacy). This pattern is consistent with the guidance provided by [13] and [212], recommending that companies should use the smallest set of short-term outcomes to create surrogate models, provided that the surrogacy assumption holds.

Interestingly, models using only one or two periods of information (where the surrogacy assumption is violated) outperform the default approach. This suggests that the benefits of noise reduction can outweigh the drawbacks of information loss. In other words, if the outcome of managerial interest (in our case, long-term cumulative purchases) has significant unexplained variations, violating the surrogacy assumption might not be the major concern. In turn, firms can improve targeting performance by using short-term outcomes in the surrogate models, even if these short-term behaviors do not capture the full impact of the intervention.

1.6 Empirical Performance: Real-world Application

This section evaluates the effectiveness of our proposed method using data from a retail-tech company in Taiwan, as detailed in Section 1.2. We begin by offering additional information about the customer covariates, followed by an empirical demonstration of the noise accumulation issue. Lastly, we showcase how our proposed solution effectively identifies the most responsive customers, thereby enhancing profitability.

1.6.1 Observed Customer Covariates

The company gathered a collection of customer covariates at the time of their initial purchase. This set included information about their first transaction, such as total sales, item count, and product categories, as well as the location of the purchase (for example, a public-access vending machine) and whether the customer was referred by a friend. Given that this data is accessible to the company prior to the execution of the intervention, it can be employed to estimate CATEs and determine whom to target. Table 1.2 presents the summary statistics of these variables for both the treatment and control groups. To maintain the company's privacy, all data points have been standardized. The results confirm successful randomization, as no significant differences exist between the groups in any of the provided variables.

Table 1.2: *Pre-treatment Covariates and Comparison across Two Experimental Conditions.*

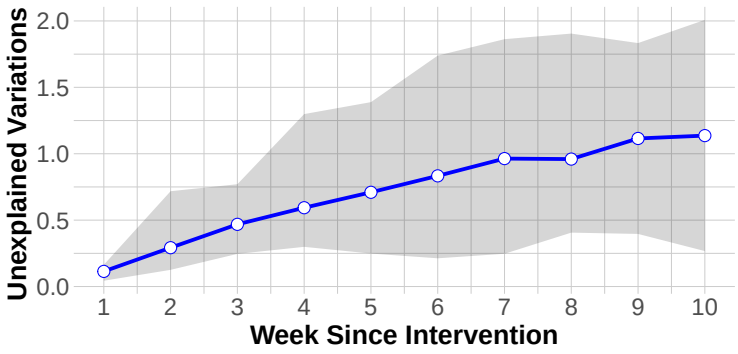
Variable	**Treatment** ($N = 889$)	**Control** ($N = 964$)	**Difference** p**-value**
log(Sales) in the first transaction	−0.0205	0.0182	0.601
log(Quantity) in the first transaction	0.0034	−0.0031	0.930
Was the first-visit fridge open to public?	0.0035	−0.0031	0.859
Did the first purchase include any side dish item?	0.0023	−0.0020	0.865
Did the first purchase include any dessert item?	0.0002	−0.0002	0.990
Did the first purchase include any beverage item?	0.0191	−0.0169	0.332
Did the first purchase include any lunch box item?	0.0128	−0.0114	0.461
Did the first purchase include any item from other categories?	0.0036	−0.0032	0.697
Was the customer referred by another customer?	−0.0016	0.0015	0.916

Note: All continuous variables were first standardized with mean zero and variance one. All binary variables were first subtracted by the mean of all customers. We use the log scale for sales and quantity to create CATE models, as outliers in these variables may impact the performance of CATE models. However, there is no significant difference for the two variables in the original scale.

1.6.2 Evidence of Noise Accumulation

As discussed in Section 1.3.1, long-term outcomes often accumulate unexplained variations which deteriorate the accuracy of CATE models, leading to suboptimal targeting. In this section, we provide empirical evidence of the noise accumulation behavior in our specific context.

First, to determine the portion of the variability in $Y_{i,T}$ that can be explained by $\mathbf{X}_i$ and W_i, we construct a regression forest model that correlates $Y_{i,T}$ with $\mathbf{X}_i$ and W_i. We then estimate the unexplained variations by calculating the absolute differences between the actual purchases and their predictions from the model. Figure 1.5 shows the unexplained variations in total purchases over T periods, ranging from $T = 1$ through $T = 10$. In the figure, the dot indicates the median, and the grey shaded region covers the top and bottom 10% of values of the distribution of unexplained variations. Notably, the unexplained variations increase as the duration of the observation period extends.

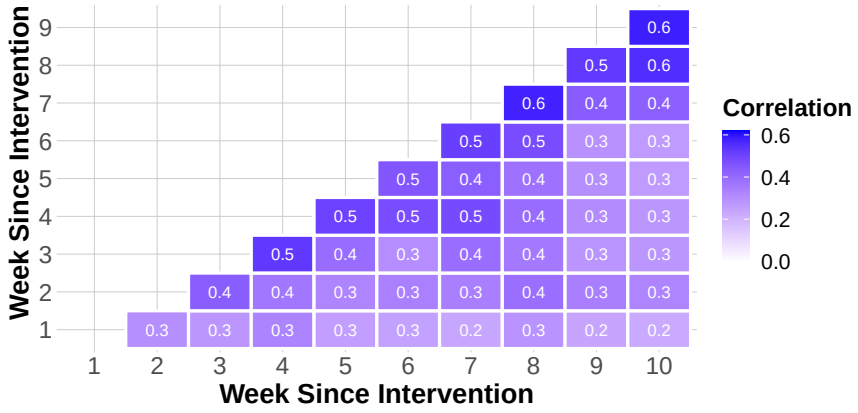


Note: Each dot illustrates the median of unexplained variations for all customers in the experimental data. The grey shaded area presents the range between the highest 10% and the lowest 10% of these variations.

Figure 1.5: *Unexplained Variations in $Y_{i,T}$ for $T = 1, \cdots, 10$.*

Next, we demonstrate the positive serial correlations of unexplained variations across different periods. Using a similar approach, we create a regression forest model, $\widehat{\mathbb{E}}[S_{i,t}|\mathbf{X}_i, W_i]$, to predict the number of purchases for each period. The residuals, $\widehat{\varepsilon}^S_{i,t} = S_{i,t} - \widehat{\mathbb{E}}[S_{i,t}|\mathbf{X}_i, W_i]$, indicate the unexplained variations in $S_{i,t}$. We then examine the cross-correlation of these residuals, $\text{Cor}(\widehat{\varepsilon}^S_{i,t_1}, \widehat{\varepsilon}^S_{i,t_2})$. Figure 1.6 presents a consistent positive correlation between the residuals for each $1 \leqslant t_1 < t_2 \leqslant 10$ highlighting a significant noise accumulation issue in our empirical context. As highlighted in Section 1.3.1, the observed positive correlation is typically attributed to unobserved

heterogeneity and attrition, both of which are highly likely to occur in this empirical context.



Note: All the correlation coefficients reported are significantly non-zero with $p < 0.01$.

Figure 1.6: *Cross-correlation Matrix of Unexplained Variations in Each Period.*

1.6.3 Empirical Analysis

Similar to our analysis in Section 1.5, we assess the effectiveness of our targeting approach against various alternatives. Specifically, we compare it to the default and myopic approaches, as well as two other imputation methods: single imputation and the BG/NBD model.[7]

We created surrogate indices using historical data from customers who were acquired at least ten weeks before the experiment began ($N = 4,031$ customers), ensuring that sufficient historical data is available to model the relationship between short- and long-term outcomes.[8] In all approaches, we use the first-week short-term outcome (i.e., $S_{i,1}$) to estimate the surrogate indices. We determine the number of short-term periods empirically. (See Appendix A.4.7 for details and complete set of results.)

[7]Unlike in simulations, we cannot substantially increase the sample size in a real-world context due to the finite number of customers the firm could acquire over time. Furthermore, we omit the results of the R-learner with lasso regularization, as it yielded identical CATE predictions for all customers. This result was expected, considering the limited sample size and the substantial unexplained variations present in the data.

[8]We use Random Forest and BG/NBD to construct those surrogate models, which are described in detail in Appendix A.4.1.

Validation Approach and Key Metrics.

To assess the targeting performance of each approach, we utilize a bootstrap validation scheme similar to that of [9]. Specifically, we generate $B = 500$ data splits consisting of training (70%) and validation (30%) sets. In each split, we estimate CATE models using the training set, with distinct outcome variables ($\ddot{Y}$) serving as the dependent variable. Then, we predict the corresponding CATEs ($\widehat{\tau}_{\ddot{Y}}$) for customers in the validation set.

Using the predictions for validation customers, we evaluate the effectiveness of each targeting approach based on two widely-used metrics: the group average treatment effects (GATEs) across predicted CATE quintile groups, and the expected profit gained by targeting customers with positive predicted CATEs.

GATEs by Predicted CATE Levels. Similar to the analyses presented in Section 1.2, we start by dividing validation customers into quintile groups based on their predicted CATEs ($\widehat{\tau}_{\ddot{Y}}$), with $\mathcal{Q}_1^{\ddot{Y}}$ having the highest predicted CATEs and $\mathcal{Q}_5^{\ddot{Y}}$ having the lowest predicted CATEs. Next, we calculate the GATE for each quintile group using the *actual* long-term outcome ($Y_{i,10}$):

$$\widehat{\text{GATE}}_{Y_{10}}(\mathcal{Q}_k^{\ddot{Y}}) = \frac{\sum_{i:\ i \in \mathcal{Q}_k^{\ddot{Y}}, W_i = 1} Y_{i,10}}{\left|\{i:\ i \in \mathcal{Q}_k^{\ddot{Y}}, W_i = 1\}\right|} - \frac{\sum_{i:\ i \in \mathcal{Q}_k^{\ddot{Y}}, W_i = 0} Y_{i,10}}{\left|\{i:\ i \in \mathcal{Q}_k^{\ddot{Y}}, W_i = 0\}\right|}.$$

Expected Profitability of Targeting Policies. To compute expected profitability, we consider a policy that targets customers with positive predicted CATEs (i.e., $\pi^{\ddot{Y}}(\mathbf{X}_i) = \mathbb{1}\{\widehat{\tau}_{\ddot{Y}}(\mathbf{X}_i) > 0\}$) and calculate the expected purchase counts in the next ten weeks using the inverse-probability-weighted (IPW) estimator [108]. Specifically,

$$\widehat{V}(\pi^{\ddot{Y}}) = \frac{1}{|\text{Validation Set}|} \cdot \sum_{i \in \text{Validation Set}} \left(\frac{\mathbb{1}[W_i = \pi^{\ddot{Y}}(\mathbf{X}_i)]}{\widehat{\mathbb{P}}[\pi^{\ddot{Y}}(\mathbf{X}_i) = W_i]} \right) Y_{i,10}, \tag{1.4}$$

where $\widehat{\mathbb{P}}[\pi^{\ddot{Y}}(\mathbf{X}_i) = W_i]$ is the (estimated) propensity score for customers who are assigned the same treatment by $\pi^{\ddot{Y}}$ as in the actual data.[9] While the treatment assignment in the data is random and independent of the derived targeting policy, we utilize IPW adjustment to account for any

[9]We estimate this quantity in each iteration using the probability forest implemented by the `grf` package.

possible imbalances between treated and non-treated customers because the sample used for profit evaluation (i.e., validation customers who were assigned the same treatment in the actual data as $\pi^{\ddot{Y}}$ assigns for policy evaluation) is relatively small. The IPW adjustment is also frequently employed in other marketing literature that utilizes randomized controlled experiments for policy evaluation [e.g., 218, 103].

We then calculate the expected profit under policy $\pi^{\ddot{Y}}$ using the following formula:

$$\text{Profit}(\pi^{\ddot{Y}}) = \text{AOV} \cdot \left[p \cdot \widehat{V}(\pi^{\ddot{Y}}) - d \cdot \left(\frac{\sum_{i=1}^{N} \pi^{\ddot{Y}}(\mathbf{X}_i)}{N} \right) \cdot U^{W=1} - d \cdot \left(1 - \frac{\sum_{i=1}^{N} \pi^{\ddot{Y}}(\mathbf{X}_i)}{N} \right) \cdot U^{W=0} \right],$$

where AOV is the average order value,[10] p is the average profit margin, $d = 15\%$ is the discount the coupon provided, U^W is the average number of coupons being used under the treatment condition W, and $\frac{\sum_{i=1}^{N} \pi^{\ddot{Y}}(\mathbf{X}_i)}{N}$ calculates the proportion of customers being treated under policy $\pi^{\ddot{Y}}$.

When assessing the profitability of different approaches, we also investigate the *policy learning* approach for determining targeting policies. As discussed in Section 1.3.2, using the actual long-term outcome for policy construction could lead to issues with noise accumulation. To address this, one could incorporate our proposed method into policy learning. This involves replacing the actual long-term outcome with a proxy that has reduced variance, and subsequently applying policy learning using this proxy as the outcome variable.

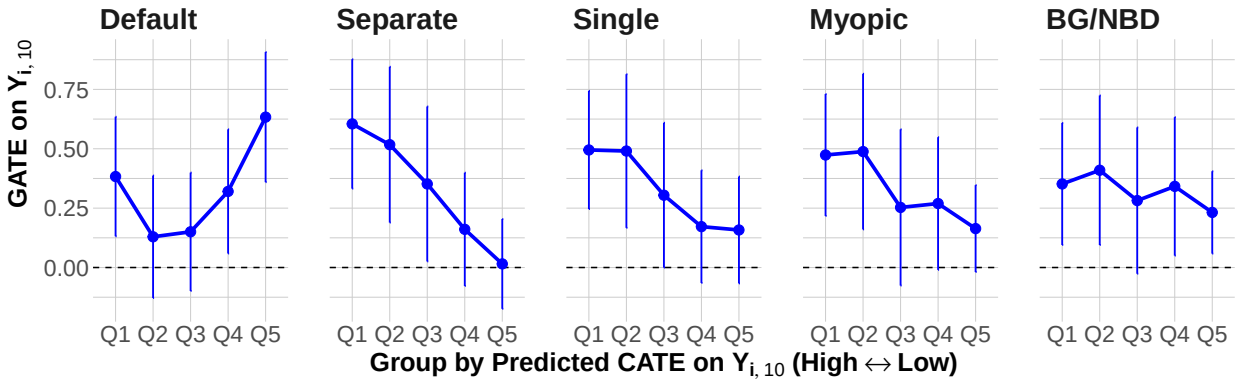
Specifically, we utilize the doubly robust policy learning technique introduced by [17]. Our method involves several steps: First, we estimate the doubly robust (DR) scores using various outcome variables ($\ddot{Y}$) as the proxy for CATE. We then develop targeting policies ($\pi_{\ddot{Y}}$) by creating cost-sensitive classifiers that predict which customers would have positive DR scores, using the DR score itself as the misclassification cost. Finally, we compute the expected profit improvement for each policy. Detailed information on the implementation is available in Appendix A.4.3.

[10] Note that we did not observe a significant difference in AOV between the treatment and control groups (Mean difference = \$0.05 with a p-value of 0.88).

1.6.4 Empirical Results

GATEs by Predicted CATE Levels.

Figure 1.7 shows the GATEs by predicted CATE groups. As discussed in Section 1.2, the U-shaped curve generated by the default approach indicates that the CATE model for $Y_{i,10}$ is unable to identify customers with the highest or lowest incremental effects. In contrast, all models that use short-term signals to estimate CATEs are more effective at ranking customers' long-term treatment effect than the default method.



Note: Each point represents the mean of bootstrap results on the validation customers together with the one standard deviation interval. Groups $\mathcal{Q}_1^{\ddot{Y}}, \cdots, \mathcal{Q}_5^{\ddot{Y}}$ are categorized based on the decreasing order of treatment effects predicted by CATE models for various outcome variables. GATEs are computed on the actual long-term outcome ($Y_{i,10}$). We present the results from T-learner as it gives the best targeting profitability. Our findings are robust across different CATE models. See Appendix A.4.5 for the results from other CATE models.

Figure 1.7: *Actual GATEs by Predicted CATE Levels for Different Outcome Variables.*

Among all these models, the separate imputation method produces the best targeting performance as it generates the steepest curve. Specifically, the GATE for $\mathcal{Q}_1^{\ddot{Y}}$ (representing the most sensitive customers identified by the method) is much larger than that of $\mathcal{Q}_2^{\ddot{Y}}$, larger than that of $\mathcal{Q}_3^{\ddot{Y}}$, and so on. Conversely, the BG/NBD model yields the least favorable result among all the proxies. This finding aligns with our intuition that the BG/NBD approach is likely to be ineffective when the parametric specifications of key parameters are different from the actual relationships, which is likely to be the case in this empirical application.

Profitability of Targeting Policies.

We compare the expected profit of each targeting policy, denoted by $\pi^{\ddot{Y}}$ (as described in Section 1.6.3), with the profit the company would obtain if it ran a uniform policy, π_0, which applies the best intervention uniformly to all customers. In our scenario, given that the average treatment effect is positive, π_0 involves sending three coupons to every customer. Specifically, we define the profit improvement (PI) for each targeting approach as: $\text{PI}(\pi^{\ddot{Y}}) = \frac{\text{Profit}(\pi^{\ddot{Y}})}{\text{Profit}(\pi_0)} - 1$, where $\text{Profit}(\pi_0)$ is calculated in the same way as $\text{Profit}(\pi^{\ddot{Y}})$.

Table 1.3 presents the results of different approaches. The first column, labeled "Predicted CATEs," illustrates the profit improvement achieved when targeting rules are based on predicted CATEs.[11] The second column, labeled "Policy Learning," details the profit improvement attained by utilizing the doubly robust policy learning approach with various outcome variables.

Table 1.3: *Expected Profit Improvement: Targeting Based on Predicted CATEs or Targeting Using Policy Learning Based on Different Outcome Variables.*

Outcome Variable $\ddot{Y}$	**Predicted CATEs**	**Policy Learning**
Separate Imputation	5.81% (4.19%)	4.57% (3.90%)
Single Imputation	4.06% (4.51%)	3.66% (4.75%)
BG/NBD Imputation	0.95% (2.77%)	−0.26% (4.36%)
Myopic	3.34% (3.60%)	3.51% (4.58%)
Default	−3.52% (3.17%)	−3.44% (3.37%)

Note: We average the profit improvement over 200 replications and show in parentheses the standard deviations. For the "Predicted CATEs" approaches, we present the results from T-learner as it gives the best targeting profitability. Our findings are robust to using different CATE models.

Several key results are worth highlighting. Firstly, consistent with our simulation results, the separate imputation approach (shown in the first row) consistently delivers the highest expected profits, whether we use predicted CATEs or policy learning for targeting. In contrast, the default method (last row) consistently leads to a loss in profit. This suggests that basing targeting decisions on predicted CATEs derived from actual long-term outcomes can detrimentally affect the long-term profitability, highlighting the substantial impact of noise accumulation. Interestingly, the myopic approach (fourth row) also mitigates much of the profit loss caused by high noise levels. This suggests that in certain scenarios, using less information — such as fewer observed

[11]We present the results when using a T-learner to compute CATEs. For robustness, results from alternative CATE estimation methods are provided in Appendix A.4.6, which show consistent findings.

periods — can, paradoxically, lead to better outcomes for long-term profitability.

Secondly, the separate and single imputation methods (first and second rows, respectively) outperform the myopic strategy in terms of profit. This highlights the importance of connecting short-term and long-term outcomes. However, the BG/NBD model (third row) falls short in performance compared to other short-term proxies, suggesting that flexible machine learning methods may be more advantageous in our empirical context. Lastly, the results obtained from policy learning are marginally less effective than those based on predicted CATEs, potentially due to the lower accuracy of the doubly robust scores compared to the predicted CATEs from the CATE model.

1.7 Conclusion and Future Directions

Firms often leverage targeted interventions to improve their business outcomes. An increasingly popular approach is to combine experimentation (or A/B testing) and customer data to predict each customer's sensitivity to an intervention. However, our research — both from theoretical and empirical perspectives — shows that this method can be ineffective in situations where outcomes are the result of recurrent behaviors that accumulate unexplained variations over time. To address this challenge, we propose a new targeting strategy that emphasizes *reducing the noise in outcome variables* prior to estimating CATE models. This method enhances the precision in estimating customers' long-term sensitivity to interventions, thereby significantly improving targeting efficacy.

Specifically, our proposed solution involves developing a *surrogate index* using *separate imputation* to effectively address the challenge of estimating long-term CATE. This method utilizes short-term behavioral changes to infer long-term treatment effects, while distinctively accounting for the dynamics of customer attrition and purchase intensity. As a result, it effectively reduces the impact of unexplained variations when estimating CATE for long-term outcomes, offering significant performance improvements over current methodologies. Our solution is readily applicable using commonly accessible machine learning algorithms, rendering it practical for a broad range of businesses, spanning industries with both contractual and non-contractual customer relationships. By capitalizing on their existing historical purchase data, companies can improve

their targeting strategies without the need for expanding the size of their experiments, thus avoiding extra costs.

We evaluate our proposed solution using both simulation analyses and a real-world marketing campaign, demonstrating superior targeting performance compared to existing practices. Our results also highlight the trade-off between information gain and noise accumulation, emphasizing the importance of balancing these factors when determining the optimal number of short-term outcomes to include in a surrogate index model. Our findings indicate that when the long-term outcome is notably noisy, using a smaller set of short-term outcomes can outperform targeting strategies based on predicted CATEs of the actual long-term outcome, even in scenarios where the short-term outcomes do not entirely explain the long-term treatment effect. In practice, companies can conduct empirical testing to determine the most effective number of short-term outcome periods for inclusion in their surrogate models, thus maximizing their targeting performance.

While our research provides valuable insights and solutions, there are limitations that suggest directions for future research. Firstly, our proposed solution directly addresses the issue of *unobserved heterogeneity* in customer attrition and purchase intensity, which is prevalent in various marketing contexts [e.g., 72, 10]. However, other dynamics can cause more unexplained variations in the outcome variable, such as customer inertia and variety-seeking [25], state dependence [171, 59], or consumer learning [66]. Incorporating these behaviors explicitly into surrogate models may further mitigate unexplained variations and enhance targeting performance. Furthermore, there are different modeling approaches available to connect the relationship between short-term and long-term outcomes, especially when we have multiple points in time for interventions. For example, [146] proposes an innovative reinforcement learning framework that optimizes long-term customer engagement by combining immediate rewards with an estimate of residual value derived from future product usage. Thus, future research could explore the integration of these dynamics and develop new modeling approaches for surrogate index construction to enhance targeting performance.

Secondly, when the long-term outcome is a repeated purchase measure, it is natural to use short-term purchases after the intervention for surrogate index construction. However, when firms have different long-term objectives, there may not exist obvious short-term signals to use as surrogates. Therefore, it is essential to develop a general surrogate selection procedure and

document potential surrogate outcomes for various marketing applications. For example, [96] proposes an estimation method to quantify the percentage of the long-term treatment effect that short-term surrogates can explain. Additionally, [218] provides evidence that short-term conversion on subscription can be an effective low-variance proxy for long-run revenue, and [202] documents potential surrogate outcomes for the long-term user experience in the context of content recommendation. Future research could focus on identifying appropriate surrogate outcomes for different marketing contexts and developing methods to evaluate the effectiveness of these surrogates in improving targeting performance.

Thirdly, there may be scenarios where no surrogate outcome is available for noise reduction, such as when the objective is to directly optimize a short-term outcome with significant unexplained variations. In such cases, future research could explore the development of new CATE models that are more resilient to noise in the outcome variable. For example, [111] proposes an iterative error correction procedure to improve CATE estimation when the data is intentionally masked by large noise for privacy protection. Furthermore, it would be worthwhile to investigate how to incorporate the estimation uncertainty of CATEs into the targeting strategy and determine whether it can further enhance the profitability of a marketing campaign.

Finally, the proposed imputation strategy relies on state-of-the-art machine learning methods to predict future purchases based on observed short-term behaviors. However, machine learning models may also overfit large unexplained variations in historical data, resulting in inaccurate long-term outcome predictions. Future research could explore alternative imputation strategies that are more robust to data noise. For instance, [158] proposes a Bayesian approach to predict purchase likelihood by incorporating information from intermediate stages in the customer journey. It would be worthwhile to investigate whether their approach can further mitigate the impact of unexplained variations in historical data.

Chapter 2

Enhancing Treatment Effect Prediction on Privacy-Protected Data: An Honest Post-Processing Approach[1]

2.1 Introduction

As companies increasingly rely on customer data for personalization, privacy concerns have grown among firms, regulators, and consumers. Traditional methods like anonymization and hashing are widely used but have proven vulnerable to re-identification attacks [e.g., 190, 150, 2, 47]. The European Data Protection Board (EDPB) [70] and the U.S. Federal Trade Commission [76] warn that these techniques do not fully prevent re-identification and may expose organizations to legal risks.

To address this, Differential Privacy (DP) [60] has emerged as a fundamental privacy-preserving framework, ensuring that analyses do not expose sensitive individual information by injecting noise into the data before release. In particular, Local Differential Privacy (LDP) [124] provides a decentralized approach by applying noise directly to individual data points before they are collected or stored, ensuring that even the data curator only accesses privatized information. This makes LDP particularly valuable for aggregating data from decentralized sources or scenarios

[1] Coauthored with Eva Ascarza.

where data privacy must be ensured even against the data collector, reducing risks associated with breaches or unauthorized access. Consequently, LDP has been widely adopted across industries, including internet services [68], digital advertising [82, 195], mobile applications [**?**], and IoT devices [198]. It has also been a key focus in emerging data privacy guidelines, such as those from the U.S. National Institute of Standards and Technology [151].

To illustrate how LDP protects organizations, consider an e-learning platform selecting students for a promotional offer based on demographic and course records. To opfutimize the offer, the platform runs an experiment where a randomly selected group receives emails, while others do not. To reduce privacy risk, the dataset may be anonymized by removing direct identifiers such as names and emails. However, privacy attackers with access to both anonymized data — whether through insider threats or data breaches — and public LinkedIn profiles could still re-identify students by matching shared attributes such as age, education, and course certificates. Such re-identification risks can undermine user trust, expose the platform to regulatory penalties, and discourage students from sharing their course certificates. To mitigate such risk, the platform could implement LDP by adding noise to age data and randomly perturbing course records. By introducing such noise at the data collection stage, LDP significantly reduces the likelihood of successful re-identification, strengthening privacy protections while allowing the platform to collect essential data for personalization.

Although LDP is effective in preserving privacy, it presents significant challenges due to potential accuracy loss. Its core mechanism involves injecting random noise into individual-level data before collection or storage. This noise is crucial for protecting privacy but it can obscure important information, such as demographic attributes, behavioral patterns, and past interactions — factors critical for personalization. As a result, privacy-enhancing measures may inadvertently reduce personalization precision, leading to suboptimal interventions and inefficient resource allocation.

This paper examines the impact of LDP protection on a key personalization practice: designing targeted interventions through Conditional Average Treatment Effect (CATE) estimation [e.g., 114, 9, 98, 104, 189]. CATE quantifies the expected difference in outcomes between treated and untreated individuals with similar characteristics, serving as a fundamental metric for determining whom to treat. Decision-makers often leverage randomized controlled experiments to estimate

CATE and identify individuals most likely to benefit from an intervention, enabling highly personalized strategies. In marketing, for example, CATE estimation helps firms prioritize which customers should receive a promotional offer by identifying those most likely to increase their purchases in response [e.g., 9, 137, 184, 63]. This targeted approach allows companies to allocate resources more efficiently by focusing on customers with the highest incremental response, thereby maximizing return on investment. However, LDP-induced noise distorts individual characteristics and weakens the ability to capture treatment-response relationships, compromising both prediction accuracy and the effectiveness of treatment prioritization.

The objectives of this research are twofold. First, we examine how LDP affects the performance of state-of-the-art CATE models. We show that LDP implementation can significantly distort both bias and variance, resulting in model-dependent distortions that vary across covariates. These distortions have critical implications for decision-makers, as they can lead to resource misallocation when determining who should receive an intervention. Beyond illustrating the privacy-utility trade-off, our findings suggest that existing correction methods from the measurement error literature are insufficient. Traditional approaches often rely on model-specific assumptions, struggle to correct biases in high-dimensional settings, and fail to address variance inflation introduced by LDP noise. Moreover, many of these methods require the collection of additional clean data, which contradicts the privacy-preserving objectives of LDP. These challenges highlight the need for a generalizable, model-agnostic approach to mitigate LDP's impact on CATE prediction while maintaining privacy guarantees.

Second, we propose a *post-processing* approach that improves the predictive accuracy of CATE models while preserving privacy guarantees. Our method constructs a sequence of *adjustment models*, each iteratively refining the CATE predictions by aligning them more closely with an unbiased but noisy proxy for the true CATE. In each iteration, we update the CATE estimates by adding the adjustment model, scaled by a *step size* that optimally balances accuracy improvements. Specifically, the adjustment model predicts the residual between the current CATE estimate and the proxy CATE, while the step size is chosen by minimizing the sum of squared residuals to ensure an optimal update magnitude. This iterative process continues until further reductions in residuals are no longer possible. Unlike traditional measurement error correction techniques, the post-processing approach provides a privacy-preserving, model-agnostic solution that improves

predictive accuracy without requiring additional clean data or strong modeling assumptions.

A key challenge in applying post-processing to CATE estimation is that this approach relies on a *noisy proxy* for the true treatment effect, as the actual CATE is never directly observed. Aligning predictions with this imperfect proxy can introduce overfitting, reducing the model's ability to generalize to new individuals. To address this, we propose an *honest model adjustment* approach. We split the (privacy-protected) data into three disjoint sets: one for training the initial CATE model, another for post-processing, and a third for determining when to stop adjustments. This sample-splitting ensures that aligning predictions with the noisy proxy improves accuracy as it prevents overfitting that would occur if the same data were reused. Furthermore, we introduce *subgroup cross-learning*: Instead of updating the model based on the entire adjustment dataset, we partition it into subgroups. Each iteration trains the adjustment model on one subgroup while using the remaining subgroups to determine the optimal update step. This method decouples adjustment model learning from step-size determination, thereby ensuring honesty in post-processing and mitigating overfitting caused by noisy proxy estimates.

Through extensive analyses of both simulated and real-world datasets, we provide strong evidence that our method enhances effectiveness under LDP protection. By evaluating scenarios where covariates and the outcome variable are protected by LDP, we demonstrate that our approach significantly improves CATE accuracy and leads to more effective treatment prioritization. Our method offers several practical advantages for firms. First, it is privacy-preserving, maintaining LDP guarantees without requiring data denoising or access to clean data. Second, it is cost-efficient, enhancing accuracy without the need for large experimental datasets while keeping post-processing computationally manageable. Third, it is easy to implement, integrating seamlessly with standard machine learning models and existing software packages, enabling quick and scalable deployment.

Our contributions are twofold. Substantively, we provide comprehensive evidence — through theoretical analysis, simulations, and empirical applications — on how LDP affects CATE prediction and treatment prioritization. To our knowledge, this is the first study to examine this issue in depth. We show that LDP-induced noise can lead to significant efficiency loss, highlighting the need for systematic evaluation of privacy–utility trade-offs across data-driven marketing practices. Methodologically, we propose a novel *honest post-processing* approach that improves CATE

prediction accuracy under LDP protection and, more broadly, in noisy data settings. It provides decision-makers with a practical way to mitigate accuracy loss when privacy mechanisms are in place, and supports better treatment prioritization without compromising individual-level privacy. We position post-processing as a promising strategy for improving model accuracy in high-noise settings — whether due to LDP protection or traditional measurement error — highlighting both the practical relevance and methodological contributions of our work.

This paper is structured as follows. Section 2.2 outlines the connections to existing literature. Section 2.3 characterized the impact of LDP on popular CATE models. Section 2.4 introduces the proposed solution and discusses its benefits. We evaluate the empirical performance of the proposed solution using simulation studies in Section 2.5 and with two real-world datasets in Section 2.6. Finally, we conclude in Section 2.7 with recommendations for future research.

2.2 Related Literature

Our research connects to multiple streams of literature across marketing, statistics, economics, and computer science.

2.2.1 CATE Estimation and Targeted Interventions

Our research contributes to the growing literature on CATE estimation and targeted interventions. Several approaches have been proposed to estimate CATEs using experimental data, which can be broadly classified into three types [104]. *Direct methods*, such as Causal Forest [199], estimate treatment effects by optimizing a loss function specifically designed for capturing treatment effect heterogeneity. *Transformed outcome regression* methods [e.g., 155, 180, 126] construct a statistical proxy for the true CATE — such as Robinson's transformation used in R-learners [155] — and then train a CATE model to predict this proxy using observed covariates. *Indirect methods*, such as the T-learner [134], estimate outcomes separately for the treated and untreated groups using observed covariates, and then compute the difference between these predictions to obtain the treatment effect estimate. Each method has its strengths and is suited to different data characteristics and practical applications [170]. Our research extends this literature by introducing a model-agnostic framework that improves predictive accuracy for all these models, particularly in settings where

data contains substantial noise.

The application of CATE models in marketing has become increasingly relevant across various contexts, including customer retention [9, 137, 212], membership subscription [184, 217], pricing and promotions [e.g., 186, 63], and catalog mailing campaigns [104]. A growing stream of research examines the challenges that CATE models face in low signal-to-noise marketing environments. For example, [112] show that the variance of common CATE models increases with outcome noise and propose leveraging low-variance signals to mitigate this issue. Our work extends this research in two ways. First, we broaden the theoretical analysis to cases where noise affects both the outcome variable and individual covariates. Second, we introduce a new approach for improving CATE prediction in high-noise environments that does not depend on low-variance signals.

2.2.2 Classical Measurement Error

Our research contributes to the literature on classical measurement errors, which arise in our context when decision-makers introduce noise into individual data for privacy protection. Prior studies [e.g., 43, 24, 1] demonstrate that such errors can bias regression coefficients and average treatment effect estimates in observational settings. We extend this line of research by conducting a comprehensive theoretical and empirical analysis of how measurement errors affect both the bias and variance of CATE predictions when using flexible machine learning models.

Existing bias correction methods for classical measurement errors typically follow four approaches: utilizing repeated measurements of variables [e.g., 99, 138, 175, 3], incorporating additional clean data [e.g., 181, 136, 38, 109], applying instrumental variables [e.g., 99, 176, 110, 213], and leveraging noise distribution information [e.g., 159, 207, 34, 67, 33, 177]. However, these methods have significant limitations when correcting errors in CATE predictions under LDP protection. First, they often assume strong parametric restrictions, design for low-dimensional covariates, and are tailored to specific model classes, limiting their applicability (see Section 2.4.1 for a detailed discussion) for CATE prediction. Second, many existing approaches rely on clean or denoised data, which conflicts with privacy preservation. Our research addresses these challenges by proposing a model-agnostic solution that improves CATE prediction without requiring clean data, instrumental variables, or repeated measurements– making it broadly applicable for enhancing

personalized interventions under LDP protection.

2.2.3 Local Differential Privacy

Given its growing importance in real-world applications, LDP has been the focus of an expanding body of research. To date, most LDP research has concentrated on developing new privacy-protection mechanisms and establishing their theoretical privacy guarantees [e.g., 124, 68, 56]. More recently, studies like [156] have explored private CATE estimation within the centralized differential privacy framework. Other research examines the privacy-accuracy trade-off, either from an information-theoretic perspective [174, 122, 225] or through a statistical accuracy framework [183, 6]. Despite the widespread adoption of LDP by major tech companies, its implications for CATE prediction and targeted interventions remains largely unexplored. This research takes a step forward in understanding how LDP affects this critical practice.

2.2.4 Model Calibration

Methodologically, our research builds on the literature on calibration — a post-processing technique that adjusts a model's predictions so that its predicted probabilities accurately reflect the true likelihood of the positive event, conditional on the predicted probability [141, 163]. Recently, the model calibration framework has been extended to CATE estimation. [206] formally defines calibration in the context of CATE estimation and derives a proper loss function to assess whether a CATE model is well-calibrated. Meanwhile, [127] apply the model calibration framework to enhance the generalizability of CATE models for new individuals with significantly different covariate distribution.

Our work makes several contributions to this literature. Conceptually, we demonstrate that the idea of adjusting predictions within subgroups can be extended to address high-noise challenges in CATE estimation, such as LDP protection and measurement error. This expands the scope of existing literature, which has primarily focused on improving fairness across groups [e.g., 100, 32, 83], enhancing generalizability under covariate shifts [130, 127], or optimizing prediction performance across multiple loss functions [88].

The proposed algorithm introduces several novel elements. First, while the idea of correcting

predictions within subgroups originates from [219], our approach is unique in that it leverages data from other subgroups to determine the magnitude of updates, ensuring overall accuracy, a technique not previously explored in the literature. Second, while existing research [206] emphasizes the importance of using separate datasets to estimate nuisance components (e.g., propensity scores or conditional mean outcome models) when constructing calibration targets for post-processing CATE predictions, it does not highlight the need to separate the data used for initial model training from that used in post-processing. We theoretically justify and empirically validate the importance of this data-splitting strategy, which we refer to as the honesty principle in post-processing.

2.3 Problem: Impact of LDP Protection on CATE Prediction

We begin by examining the impact of LDP on CATE prediction. This analysis serves two key purposes. First, it provides insight into how LDP protection affects the bias and variance properties of commonly used CATE models. Second, it uncovers critical characteristics of the prediction error that inform the design of an effective solution (Section 2.4).

2.3.1 Problem Formulation and Data Collection

Consider a decision-maker who wants to improve a specific outcome (Y_i) through an intervention with two treatment conditions ($W_i \in \{0, 1\}$). The decision-maker hypothesizes that the intervention could potentially alter the value of Y_i for each individual i, but the impact of this intervention may vary among individuals with different characteristics. This heterogeneous impact is characterized as the conditional average treatment effect (CATE), which measures the difference in expected outcomes among individuals based on their treatment assignment and pre-treatment characteristics. Mathematically, the CATE is defined as follows:

$$\tau(\mathbf{X}_i) \equiv \mathbb{E}[Y_i(1) \mid \mathbf{X}_i] - \mathbb{E}[Y_i(0) \mid \mathbf{X}_i],$$

where $Y_i(W_i)$ is the potential outcome [172] of individual i's response given the treatment condition W_i, and $\mathbf{X}_i$ denotes a set of pre-treatment characteristics that are believed to moderate the treatment effect.

Following common practice [9, 104], the decision-maker constructs a CATE model, $\widehat{\tau}(\cdot \mid \mathcal{D})$, using the dataset $\mathcal{D}$ to predict treatment effects for new individuals, helping to prioritize who should receive the intervention in the future. To mitigate privacy risks, the decision-maker applies LDP to protect the experimental data, resulting in a privatized dataset denoted as $\widetilde{\mathcal{D}}$. Consequently, the CATE model can only be estimated using $\widetilde{\mathcal{D}}$, which we denote as $\widehat{\tau}(\cdot \mid \widetilde{\mathcal{D}})$.

Setup.

In line with the existing literature [199, 180, 155, 126], we assume that the collected data satisfies the following assumptions.

Assumption 2.1 (Assumptions on Data Collection)

1. (Unconfoundedness) *The treatment assignment within the experiment set is independent of their potential outcomes given the true customer covariates, i.e.,* $Y_i(1), Y_i(0) \perp\!\!\!\perp W_i \mid \mathbf{X}_i$.

2. (Overlap) *The probability of an individual receiving (or not receiving) a treatment should always be positive. This can be represented as follows:* $0 < \mathbb{P}[W_i = 1 \mid \mathbf{X}_i] < 1$ *for any possible* $\mathbf{X}_i$.

3. (No Interference) *The potential outcomes of each individual are independent of the treatments received by other individuals, i.e.,* $Y_i(1), Y_i(0) \perp\!\!\!\perp W_j,\ \forall j \neq i$.

LDP Protection.

We examine a scenario in which the decision-maker applies LDP to protect individual data, including the outcome of interest and covariates while leaving the treatment status (e.g., whether an individual receives a promotion) unprotected, as it is typically neither sensitive nor useful for a privacy attacker to infer individual identity. In this case, the decision-maker would add noise to the outcome of interest, ensuring that only a noisy version ($\widetilde{Y}_i$) is observed instead of the true outcome (Y_i). Without loss of generality, we represent LDP protection as additive noise: $\widetilde{Y}_i = Y_i + \eta_i^Y$. Similarly, individual covariates are protected so that the decision-maker observes only a noisy version, $\widetilde{\mathbf{X}}_i = \mathbf{X}_i + \boldsymbol{\eta}_i^X$. In this setting, the decision-maker relies solely on the noisy experimental data $\{(\widetilde{\mathbf{X}}_i, \widetilde{Y}_i, W_i)\}_{i=1}^N$ with N samples to construct CATE models.

The additive noise assumption covers a wide range of LDP mechanisms, including the Laplace mechanism [61, 124] for continuous data and randomized response [204, 68] for categorical data. These techniques introduce noise by either perturbing numerical values or flipping categorical labels with a predefined probability, ensuring compliance with LDP definitions [208]. Since both approaches can be expressed as additive noise, our theoretical results extend to several privacy-preserving implementations.[2]

2.3.2 Characterizing Bias and Variance Caused by LDP Protection

When selecting recipients for interventions based on predicted CATEs (e.g., targeting those exceeding key thresholds or ranking among top performers), both prediction bias and variance significantly impact the outcome of interest. High bias systematically skews treatment decisions — underestimating CATEs creates costly false negatives that leave revenue on the table, while overestimation wastes resources on non-responsive individuals. Meanwhile, high variance produces unstable predictions that undermine decision consistency and cause resource misallocation [112]. Given these direct impacts on treatment prioritization, we systematically analyze how LDP-induced noise affects both bias and variance metrics to quantify the true business cost of privacy protection.

Let $\mathbf{X}_{\text{new}}$ denote the true covariates of a new individual, and $\widetilde{\mathbf{X}}_{\text{new}} = \mathbf{X}_{\text{new}} + \boldsymbol{\eta}^{\mathbf{X}}_{\text{new}}$ the LDP-protected covariates, where $\boldsymbol{\eta}^{X}_{\text{new}}$ represents the noise introduced by LDP. The mean squared error (MSE) between the predicted and true CATE for a model trained on privacy-protected data can be expressed as:

$$\mathbb{E}\left[(\widehat{\tau}(\widetilde{\mathbf{X}}_{\text{new}} \mid \widetilde{\mathcal{D}}) - \tau(\mathbf{X}_{\text{new}}))^2\right] = \underbrace{\mathbb{E}^2\left[\widehat{\tau}(\widetilde{\mathbf{X}}_{\text{new}} \mid \widetilde{\mathcal{D}}) - \tau(\mathbf{X}_{\text{new}})\right]}_{\text{Squared Bias}} + \underbrace{\text{Var}\left[\widehat{\tau}(\mathbf{X}_{\text{new}} \mid \widetilde{\mathcal{D}})\right]}_{\text{Variance}}.$$

We analyze each component separately and provide a formal characterization of these results in Appendix B.2, where we derive the bias and variance for commonly used CATE models. In the remainder of this section, we first present the intuition behind these mathematical results and

[2]The randomized response technique can be reformulated as the injection of additive noise. Consider a binary variable Y. The randomized response with flipping probability f can be expressed as $\widetilde{Y} = Y + \mathbf{1}(Y = 1)\eta_1 + \mathbf{1}(Y = 0)\eta_0$, where $-\eta_1 \sim \text{Bernoulli}(f)$ and $\eta_0 \sim \text{Bernoulli}(f)$. A similar formulation can be applied to the dummy transformation of a discrete variable with multiple possible values. Both techniques satisfy the definition of LDP and serve as fundamental building blocks for more advanced LDP mechanisms [208].

then illustrate their effects through a simple simulation setup.

Bias. Essentially, LDP noise in both covariates and the outcome variable introduces bias into CATE models. Noise in the covariates of individuals in the training set distorts CATE estimates, similar to classical measurement error bias. Noise in the outcome variable further amplifies bias and may cause over-regularization of the CATE model. To formalize this result, we decompose the bias as follows:

$$\mathbb{E}\left[\widehat{\tau}(\widetilde{\mathbf{X}}_{\text{new}} \mid \widetilde{\mathcal{D}}) - \tau(\mathbf{X}_{\text{new}})\right] = \underbrace{\mathbb{E}\left[\widehat{\tau}(\widetilde{\mathbf{X}}_{\text{new}} \mid \widetilde{\mathcal{D}}) - \widehat{\tau}(\mathbf{X}_{\text{new}} \mid \mathcal{D})\right]}_{\text{(a) Additional Bias Caused by LDP Protection}} + \underbrace{\mathbb{E}\left[\widehat{\tau}(\mathbf{X}_{\text{new}} \mid \mathcal{D}) - \tau(\mathbf{X}_{\text{new}})\right]}_{\text{(b) Bias in the Non-private Setting}}, \tag{2.1}$$

where $\widehat{\tau}(\cdot \mid \mathcal{D})$ denotes the CATE model trained on the non-private version of the experiment data. The first term, $\mathbb{E}\left[\widehat{\tau}(\widetilde{\mathbf{X}}_{\text{new}} \mid \widetilde{\mathcal{D}}) - \widehat{\tau}(\mathbf{X}_{\text{new}} \mid \mathcal{D})\right]$, represents the difference in predictions between a CATE model trained on LDP-protected data and one trained on non-private data. The second term, $\mathbb{E}\left[\widehat{\tau}(\mathbf{X}_{\text{new}} \mid \mathcal{D}) - \tau(\mathbf{X}_{\text{new}})\right]$, is the inherent bias of a CATE model (when estimated with non-private data).

To illustrate the impact of LDP on the additional bias, we apply the Taylor approximation approach from the measurement error literature [43, 24]:

$$\mathbb{E}\left[\widehat{\tau}(\widetilde{\mathbf{X}}_{\text{new}} \mid \widetilde{\mathcal{D}}) - \widehat{\tau}(\mathbf{X}_{\text{new}} \mid \mathcal{D})\right] \approx \frac{1}{2}\,\mathbb{E}\left[\partial^2_{\mathbf{X},\mathcal{D}}\,\widehat{\tau}(\mathbf{X}_{\text{new}} \mid \mathcal{D})\right]\,\mathbf{V}(\eta^X_{\text{new}}, \boldsymbol{\eta}^{\mathcal{D}}),$$

where $\partial^2_{\mathbf{X},\mathcal{D}}\,\widehat{\tau}(\mathbf{X}_{\text{new}} \mid \mathcal{D})$ represents the second-order directional derivative of the CATE model with respect to the input covariates and the training data. The term, $\mathbf{V}(\eta^X_{\text{new}}, \boldsymbol{\eta}^{\mathcal{D}})$ denotes a diagonal matrix, where each diagonal element corresponds to the variance of noise injected into the covariates (η^X_{new}) and the training data ($\boldsymbol{\eta}^{\mathcal{D}}$), while all off-diagonal elements are zero.

The Taylor approximation reveals that the additional bias introduced by LDP protection depends on two key factors. First, it is influenced by the sensitivity of the CATE model to noise perturbation, captured by $\partial^2_{\mathbf{X},\mathcal{D}}\,\widehat{\tau}(\mathbf{X}_{\text{new}} \mid \mathcal{D})$. This suggests that the bias introduced by LDP protection is model-dependent and varies across covariates, unlike classical measurement error in linear models, where researchers typically characterize it as a uniform attenuation bias [1].[3]

[3]The Taylor approximation results in [43] focus on the difference between the conditional expected outcome given the true covariates and that given the noisy covariates. In our context, this corresponds to the difference between $\tau(\widetilde{\mathbf{X}}\text{new})$ and $\tau(\mathbf{X}\text{new})$. However, their analysis does not account for the model specification or the estimation

Second, it depends on the magnitude of noise introduced by LDP, represented by $\mathbf{V}(\eta^{\mathrm{X}}_{\mathrm{new}}, \boldsymbol{\eta}^{\mathcal{D}})$. Stricter privacy guarantees increase the noise, which can significantly amplify the bias.

Variance. The impact of LDP protection on CATE variance is multifaceted. On one hand, as suggested by **(author?)** [112], noise can increase the variance of CATE estimates when considering noisy outcomes and unbiased CATE models. On the other hand, excessive noise can obscure treatment effect heterogeneity, causing the model to shrink toward the average treatment effect. This, in turn, significantly reduces variance and results in nearly constant predictions. In Appendix B.2, we formally demonstrate (i) the potential dual effect of LDP protection on CATE variance and (ii) how this effect depends on the model, varies across covariates, and is influenced by the noise level.

Overall, LDP-induced noise affects both the bias and variance of CATE estimates in complex, model-dependent ways. Bias arises as noise perturbs covariates and outcomes, distorting estimates based on the model's sensitivity to input perturbations. Variance can increase due to added uncertainty in the data, but it may also decrease unpredictably if the model fails to capture heterogeneity and instead produces near-constant predictions.

2.3.3 A Simulation Example

We now empirically illustrate these findings through a simple simulation exercise. Suppose a company conducts a randomized controlled experiment with two treatment conditions ($W_i \in \{0, 1\}$) on a population of individuals (indexed by i). Let $Y_i \in \mathbb{R}$ represent the outcome of interest, and let X_i denote the pre-treatment covariate that captures heterogeneity in treatment effects. We generate simulated experimental data with $N = 2,000$ observations using the following data-generating process:

$$\begin{aligned} Y_i &= X_i^2 + W_i \cdot \underbrace{(2 + X_i - 0.5X_i^2)}_{\text{true CATE function}} + e_i, \\ W_i &\sim \text{Ber}(0.5), \quad X_i \sim \text{Unif}(-2, 2), \quad e_i \sim \mathcal{N}(0, 1). \end{aligned} \tag{2.2}$$

For illustration, we consider two scenarios separately: covariate protection under LDP and

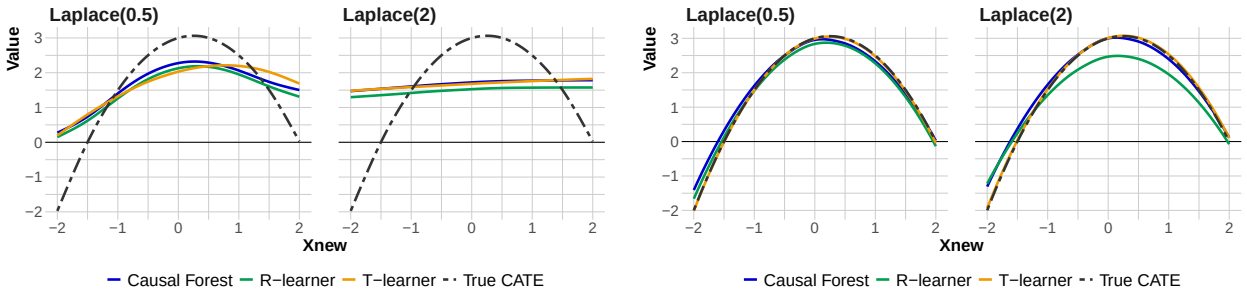
procedure used to approximate the true function, which are central to understanding the full impact of LDP in practice.

outcome protection under LDP. In the first case, when covariates are protected using the Laplace mechanism, the decision-maker observes a noisy version of the true covariates, $\widetilde{X}_i = X_i + \eta_i^X$, where η_i^X is noise drawn from a Laplace(σ) distribution. In the second case, when the outcome of interest is protected under LDP, the decision-maker observes only $\widetilde{Y}_i = Y_i + \eta_i^Y$ instead of the true outcome Y_i, with η_i^Y similarly drawn from a Laplace(σ) distribution. To assess the effect of LDP noise on CATE estimation and prediction, we generate 100 replications under each scenario and predict CATEs using three popular models: Causal Forest [199], an R-learner with XGBoost Models [155], and a T-learner with correctly specified OLS regressions[134].[4] For each covariate value X_{new}, we calculate prediction bias by averaging the predicted CATE across 100 replications and comparing it to the true CATE. To approximate variance, we compute the variance of the predicted CATE for a given covariate value across the 100 replications.

Bias. Figure 2.1 presents the bias of predicted CATEs from different models under varying noise levels. Several key patterns emerge. First, as expected, bias in predicted CATE increases with stronger privacy protection. Second, bias is model-dependent. When only the covariate is protected by LDP (Figure 2.1a), all models shrink the predicted CATE toward the average treatment effect, but the magnitude and pattern of bias vary across models. When the outcome is protected by LDP (Figure 2.1b), the differences are more noticeable, especially under high noise levels (right panel, Figure 2.1b), R-learner exhibits significantly higher bias compared to the other two approaches. Third, the direction and magnitude of bias in CATE estimation depend on the covariate values. For $X_{\text{new}} > 1.5$ and $X_{\text{new}} < -1$, all models overestimate the CATE (exacerbation bias), whereas for $-1 < X_{\text{new}} < 1.5$, they underestimate it (attenuation bias). This variation in bias direction occurs because LDP noise in CATE estimation pulls estimates toward the average treatment effect rather than toward zero.

Variance. Figure 2.2 presents the variance of CATE predictions across different covariate values and models. Similar to the results when analyzing the bias, the variance of CATE predictions is model-dependent and heterogeneous across covariate values. Moreover, when the covariate is protected by LDP (Figure 2.2a), we observe a counterintuitive result: increasing the noise level

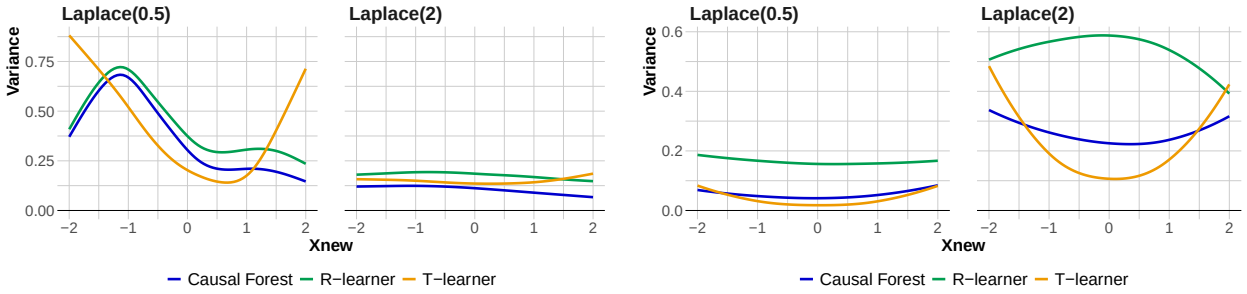
[4]For details on the model specifications, please refer to Appendix B.4.



(a) *Scenario 1: LDP-Protected Covariates* **(b)** *Scenario 2: LDP-Protected Outcome*

Figure 2.1: *Bias of CATE Predictions in the Simulation Setting*

decreases the variance. This occurs because all CATE models underestimate treatment effect heterogeneity, producing predictions that cluster around the average treatment effect. Conversely, when the outcome is protected by LDP (Figure B.1b), the variance of CATE predictions increases substantially as the noise level rises.



(a) *Scenario 1: LDP-Protected Covariates* **(b)** *Scenario 2: LDP-Protected Outcome*

Figure 2.2: *Variance of CATE Predictions in the Simulation Setting*

2.3.4 Summary

The above characterization of LDP's impact on CATE prediction highlights key considerations for developing effective mitigation methods. First, the effects of LDP-induced noise on bias and variance are strongly model-dependent, meaning different CATE models exhibit varying degrees of distortion. Second, these distortions are heterogeneous across covariates, with bias direction and magnitude varying based on covariate values. These findings highlight the necessity of a model-agnostic correction approach, one that is not tailored to any specific CATE model and one that is sufficiently flexible to adjust for prediction errors.

2.4 Solution: Honest Post-processing

2.4.1 Criteria for An Effective Solution

To mitigate LDP's impact on CATE estimation both practically and effectively, a solution must satisfy four key criteria. First and foremost, the solution must be *privacy-preserving*; it should neither rely on collecting clean, noise-free data nor involve denoising the protected data. Second, the solution must be *model-agnostic*. As companies increasingly deploy various CATE models, and the impact of LDP on accuracy varies by model, an ideal solution should not rely on model-specific approaches but should be capable of enhancing accuracy across arbitrary models. Third, the solution should be *flexible* enough to capture complex relationships and handle high-dimensional covariates. Today, decision-makers have access to high-dimensional covariates, allowing them to capture rich heterogeneity for more personalized treatment prioritization. However, this also implies complex interactions between bias, variance, and covariates across different models, making estimation and optimization more challenging. Therefore, an ideal solution must account for these relationships in high-dimensional settings to enhance accuracy across individuals. Fourth, it should address noise in both the covariates and the outcome variable.

Before introducing our proposed solution, we first assess whether existing approaches can be adapted to our needs. Table 2.1 summarizes four common strategies for mitigating measurement error bias,[5] including utilizing repeated variable measurements (row 1), incorporating additional clean data (row 2), applying instrumental variables (row 3), and leveraging noise distribution information (row 4). None of these methods fully satisfy the four key criteria outlined above. First, approaches that rely on repeated measurements or clean datasets (rows 1 and 2) are incompatible with LDP's privacy-preserving objectives, as they depend on noise-free data or denoising processes to enhance model accuracy. Second, many existing measurement error correction techniques (including those in rows 3 and 4) are tailored to specific model classes, making them difficult to generalize to state-of-the-art machine learning-based CATE models.

Third, many methods in rows 1 to 4 rely on techniques such as Generalized Method of Moments (GMM), parametric Maximum Likelihood Estimation (MLE), or deconvolution-based

[5]A more detailed review of these methods is provided in Appendix B.1.1.

Table 2.1: *Comparison of Different Solutions for Measurement Error Bias*

Solution	Overview	Privacy Protection	Model Agnostic	Flexible with Many Covariates	Related Literature
Repeated Measurements of Variables	Averages over multiple noisy measurements of the same variable to recover the true value of the variable	No	Yes	Yes	[99, 138, 175, 3]
Additional Clean Data	Utilizes a separate noise-free dataset to estimate unbiased relationships and mitigate the effects of noise in the primary data through GMM or MLE	No	Yes	No	[181, 136, 38, 109]
Instrumental Variables	Employs external instruments that are correlated with the noisy variable but uncorrelated with the noise to correct for measurement error bias	Yes	No	No	[99, 176, 110, 213]
Information of Noise Distribution	Leverages knowledge of the noise distribution to correct biases (through deconvolution or moment conditions)	Yes	No	No	[159, 207, 34, 67, 33, 177]

kernel estimators. While these approaches are theoretically well-founded, they impose strong assumptions about error distribution and functional form, which may not hold in practice. Their effectiveness further diminishes in high-dimensional settings with nonlinear and complex CATE functions. Besides, these methods come with increased computational demands and a higher risk of model misspecification, limiting their practicality in real-world applications. Finally, while they focus on correcting bias from noise in covariates, they do not address noise injected into the outcome variable, making them less applicable when the outcome is also protected by LDP.

These limitations underscore the need for new methodologies specifically tailored to improve CATE prediction under LDP protection. Such methods must be privacy-preserving, model-agnostic, scalable to high-dimensional covariates, and able to account for noise in both covariates and the outcome variable.

2.4.2 Post-processing for Enhanced CATE Prediction

Given these considerations, we propose a solution based on *post-processing*, a widely used technique in machine learning. Post-processing adjusts raw model predictions to better align with problem-specific requirements. In our context, the decision-maker first estimates a CATE model using LDP-protected data, as they typically would for learning targeting policies. This initial CATE model serves as the "raw predictions," which post-processing then refines to enhance accuracy while maintaining privacy constraints.

Post-processing has been widely used across various domains, including enhancing predic-

tive performance [77], improving model interpretability and ensuring valid inference [41], and improving transportability across different populations [130]. In our approach, post-processing refines CATE model predictions using a separate experimental dataset that is also LDP-protected but not used for training the initial CATE model. Since this dataset is similarly protected under LDP, it preserves strict privacy guarantees while refining predictions. Furthermore, because decision-makers can split the experimental dataset — using one subset to train the CATE model and another for post-processing — our approach eliminates the need for additional data collection or reliance on instrumental variables for each covariate. This makes it both practical and easily implementable in real-world settings.

Compared to standard measurement error correction methods, our post-processing approach offers several advantages for LDP-protected data. First, it does not require additional clean validation data, valid instruments, or repeated covariate measurements, aligning with privacy-preservation goals by eliminating the need for extra sensitive data. Second, because LDP guarantees remain intact even after post-processing [124], adjusting model predictions does not introduce additional privacy risks. Third, our approach is model-agnostic and flexiblle to accommodate complex relationships between prediction errors and high-dimensional covariates. Finally, it improves prediction accuracy regardless of whether the outcome or the covariates are protected by LDP.

Learning Target.

A key step in any post-processing approach is constructing an adjustment model, $\widehat{c}^{[r]}(\cdot \mid \widetilde{\mathcal{D}})$, to refine the original model's predictions. Specifically, the adjustment model estimates the difference between the target variable (e.g., the true CATE) and the model's predictions (e.g., the predicted CATE) and adjusts the model's output accordingly to compensate for this discrepancy. While constructing an adjustment model is straightforward in standard supervised learning tasks where the target variable is directly observed, applying it to CATE prediction is more challenging because the true treatment effect for each individual is fundamentally unobservable [172]. As a result, we must rely on proxy variables to approximate the true CATE, enabling us to estimate and correct the prediction error of the CATE model.

Building on recent advancements in CATE estimation [126, 206], we use the augmented inverse-propensity-weighted (AIPW) score [168] to approximate the CATE for an individual i:

$$\check{\tau}_i = \mu_1(\mathbf{X}_i) - \mu_0(\mathbf{X}_i) + \frac{W_i}{e(\mathbf{X}_i)} \left[Y_i - \mu_1(\mathbf{X}_i)\right] - \frac{1 - W_i}{1 - e(\mathbf{X}_i)} \left[Y_i - \mu_0(\mathbf{X}_i)\right],$$

where $\mu_w(\mathbf{X}_i) = \mathbb{E}[Y_i \mid W_i = w, \mathbf{X}_i]$ represents the conditional mean outcome given the treatment condition $w \in \{0, 1\}$ and the true covariates $\mathbf{X}_i$, and $e(\mathbf{X}_i) = \mathbb{P}[W_i = 1 \mid \mathbf{X}_i]$ is the propensity score of being treated given the true covariates. In practice, all the nuisance models can be estimated using cross-fitting [155].

The AIPW score is a valuable proxy due to its double robustness, ensuring it remains unbiased as long as either the conditional outcome model or the propensity score model is correctly specified. However, in our setting, we compute the AIPW score using LDP-protected data, $\widetilde{\mathcal{D}} = \{(\widetilde{Y}_i, \widetilde{\mathbf{X}}_i, W_i)\}$. This raises a potential issue: if LDP distorts *both* the conditional mean outcome model and the propensity score model, then $\check{\tau}_i$ may also be a biased approximation of the true CATE.

This concern can be resolved if decision-makers have access to the *true propensity score*. In which case, the double robustness property of the AIPW estimator [168] ensures that $\check{\tau}_i$ remains an unbiased estimator of the true CATE, even when derived using LDP-protected data. The following proposition formally establishes that the AIPW score remains an unbiased estimator of the true CATE under LDP protection when computed with the true propensity score.

Proposition 2.1 *When Assumption 2.1 holds, the AIPW score remains an unbiased proxy for the true CATE when the LDP noise has mean zero, i.e.,* $\mathbb{E}\left[\check{\tau}_i\right] = \tau(\mathbf{X}_i)$*, as long as it is computed using the true propensity score, even when both covariates and outcomes are affected by LDP noise.*

The proof is provivded in Appendix B.3. Thus, beyond Assumption 2.1 (Section 2.3.1), we assume decision-makers know the true propensity score to ensure the AIPW score's validity as a learning target. This assumption aligns well with many practical applications, as organizations frequently conduct randomized controlled experiments (i.e., A/B tests) or apply predefined business rules for treatment assignments. For example, e-commerce companies often implement rule-based treatment strategies, such as offering discounts to 80% of customers who have abandoned their carts. Because the treatment rule explicitly determines which customers receive the discounts, the true propensity score is precisely known.

Consequently, despite the noise introduced by LDP, the AIPW score remains a valid and unbiased estimator of the true CATE under reasonable and commonly satisfied conditions, making it a reliable learning target for our post-processing approach.

Post-processing Procedure.

Next, we describe how the existing post-processing framework from the machine learning literature can be used to adjust CATE predictions. Assume the decision-maker has estimated a CATE model, $\widehat{\tau}^{[0]}(\cdot \mid \widetilde{\mathcal{D}})$, using LDP-protected data $\widetilde{\mathcal{D}}$. To improve its accuracy, we aim to construct a series of *adjustment* models, $\widehat{c}^{[1]}, \cdots, \widehat{c}^{[R]}$, which iteratively refine $\widehat{\tau}^{[0]}$ and improve its predictive accuracy. In each iteration r, the CATE model $\widehat{\tau}^{[r]}$ is updated as follows:

$$\widehat{\tau}^{[r]}(\cdot \mid \widetilde{\mathcal{D}}) = \widehat{\tau}^{[r-1]}(\cdot \mid \widetilde{\mathcal{D}}) + \rho^{[r]}\widehat{c}^{[r]}(\cdot \mid \widetilde{\mathcal{D}}), \quad r = 1, \ldots, R, \tag{2.3}$$

where $\widehat{c}^{[r]}(\cdot \mid \widetilde{\mathcal{D}})$ is the adjustment model at iteration r, and $\rho^{[r]}$ is the step size used to adjust the predictions. Here, our objective is to ensure that, in each iteration, the CATE predictions are adjusted to better align with the AIPW score by minimizing the sum of squared (proxy) residuals over the (LDP-protected) experiment data:

$$\sum_{i \in \widetilde{\mathcal{D}}} \left[\widehat{\tau}^{[r-1]}(\cdot \mid \widetilde{\mathcal{D}}) + \rho^{[r]}\widehat{c}^{[r]}(\cdot \mid \widetilde{\mathcal{D}}) - \check{\tau}_i \right]^2. \tag{2.4}$$

We apply *gradient boosting* [77] to solve this optimization problem. Specifically, in each iteration r, we first construct an adjustment model $\widehat{c}^{[r]}$ that predicts the direction in which the CATE model should be adjusted to reduce the squared proxy residuals. This direction corresponds to the negative gradient of the objective function (2.4). Mathematically, the negative gradient for each individual is given by:

$$-\frac{\partial(\widehat{\tau}^{[r-1]}(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}) - \check{\tau}_i)^2}{\partial \widehat{\tau}^{[r-1]}(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}})} = 2\left[\check{\tau}_i - \widehat{\tau}^{[r-1]}(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}})\right].$$

Thus, the adjustment model is trained to predict $\check{\tau}_i - \widehat{\tau}^{[r]}(\widetilde{\mathbf{X}}_i)$ using $\widetilde{\mathbf{X}}_i$ as input variables.

Once the adjustment model is learned, we determine the optimal step size $\rho^{[r]}$ to control the

magnitude of the update. This is done by solving the following minimization problem:

$$\rho^{[r]} = \arg\min_{\rho} \sum_{j \in \widetilde{\mathcal{D}}} \left[\widehat{\tau}^{[r-1]}(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}) + \rho \widehat{c}^{[r]}(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}) - \check{\tau}_i \right]^2. \tag{2.5}$$

This step intends to minimize the squared proxy residual by determining the appropriate adjustment magnitude, $\rho^{[r]}\widehat{c}^{[r]}(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}})$, to apply to the CATE model.

Finally, this two-step process is repeated iteratively until convergence, i.e., when further reducing the sum of squared residuals is no longer possible.

2.4.3 Challenges of Implementing Post-Processing in the Context of LDP

While $\check{\tau}_i$ provides an unbiased signal for the true treatment effects (Proposition 2.1), it remains a noisy proxy under LDP protection for two key reasons: (i) the LDP-protected outcome $\widetilde{Y}_i$ includes injected noise, reducing its precision of the AIPW score as an estimate for the true CATE, and (ii) the nuisance models $\widehat{\mu}_w^{\widetilde{Y}}(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}})$ are less accurate when estimated using LDP-protected data. Although Proposition 2.1suggests that post-processing should not introduce bias, the presence of LDP noise can still significantly impact the effectiveness of standard post-processing methods. Specifically, LDP noise increases the approximation error of the AIPW score as a proxy for the true CATE, which raises the risk of overfitting in standard post-processing methods and can lead to higher variance in the final predictions.

To formally illustrate the overfitting risk when aligning the current CATE model to the AIPW score, we decompose the expected squared proxy residual as follows:

$$\begin{aligned}
\mathbb{E}_{\widetilde{\mathcal{D}}}\left[(\widehat{\tau}^{[r]}(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}) - \check{\tau}_i)^2\right] &= \mathbb{E}_{\widetilde{\mathcal{D}}}[(\widehat{\tau}^{[r]}(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}) - \tau(\mathbf{X}_i) - \underbrace{(\check{\tau}_i - \tau(\mathbf{X}_i))}_{\equiv \check{\epsilon}_i})^2] \\
&= \underbrace{\mathbb{E}_{\widetilde{\mathcal{D}}}\left[(\widehat{\tau}^{[r]}(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}) - \tau(\mathbf{X}_i))^2\right]}_{\text{True mean squared prediction error}} - 2 \cdot \underbrace{\mathbb{E}_{\widetilde{\mathcal{D}}}\left[(\widehat{\tau}^{[r]}(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}) - \tau(\mathbf{X}_i)) \cdot \check{\epsilon}_i\right]}_{\text{Comovement w.r.t. the proxy error}} + \underbrace{\mathbb{E}_{\widetilde{\mathcal{D}}}\left[\check{\epsilon}_i^2\right]}_{\text{Variance}},
\end{aligned} \tag{2.6}$$

where $\check{\epsilon}_i \equiv \check{\tau}_i - \tau(\mathbf{X}_i)$ denotes the approximation error of the AIPW score. The first term, $\mathbb{E}_{\widetilde{\mathcal{D}}}\left[(\widehat{\tau}^{[r]}(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}) - \tau(\mathbf{X}_i))^2\right]$, is the true mean squared prediction error that the post-processing procedure seeks to minimize. The second term, $\mathbb{E}_{\widetilde{\mathcal{D}}}\left[(\widehat{\tau}^{[r]}(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}) - \tau(\mathbf{X}_i)) \cdot \check{\epsilon}_i\right]$, captures the co-movement between the model prediction error and the approximation error of the AIPW score. The third term, $\mathbb{E}_{\widetilde{\mathcal{D}}}\left[\check{\epsilon}_i^2\right]$, represents the variance of the AIPW score and remains unchanged

throughout the post-processing procedure.

This decomposition highlights the potential for overfitting when refining CATE predictions using the AIPW score in post-processing. At first glance, minimizing the squared proxy residual on $\widetilde{\mathcal{D}}$ may seem to improve CATE prediction. However, this approach does not necessarily reduce the true prediction error. Instead, it risks aligning the post-processed CATE model $\widehat{\tau}^{[r]}$ too closely with the approximation error $\check{\epsilon}_i$, which arises from noise in the AIPW score.

To understand why, consider the first and second terms in Equation (2.6). When minimizing squared proxy residuals, a standard post-processing method cannot distinguish between reducing the true mean squared prediction error (first term) and inadvertently fitting to the approximation error $\check{\epsilon}_i$, thereby increasing the second term's magnitude. As a result, the post-processed CATE model partially captures noise by overfitting $\check{\epsilon}_i$, leading to higher variance in final predictions and potential degradation in overall model accuracy.

To tackle this challenge, we propose the *honesty* principle in post-processing to mitigate overfitting. Specifically, whenever we build an adjustment model $\widehat{c}^{[r]}$, determine the step size $\rho^{[r]}$, or evaluate performance improvements, we propose to use a *separate* dataset that was not used in the construction of previous models. We refer to this approach as *honest* because it adheres to the honesty principle used in Causal Trees [14], where the data is partitioned into two disjoint sets — one for model construction and another for generating predictions.

To understand how honesty helps mitigate the overfitting problem characterized in (2.6), let us consider two *disjoint* datasets: $\widetilde{\mathcal{D}}_1$, used for constructing the CATE model $\widehat{\tau}^{[r]}(\cdot \mid \widetilde{\mathcal{D}}_1)$, and $\widetilde{\mathcal{D}}_2$, reserved for post-processing $\widehat{\tau}^{[r]}$ by aligning its predictions with the AIPW score of individuals in $\widetilde{\mathcal{D}}_2$. By ensuring statistically independence between $\widetilde{\mathcal{D}}_1$ and $\widetilde{\mathcal{D}}_2$, we can express the expected squared proxy residual for individual j in $\widetilde{\mathcal{D}}_2$ as follows:

$$\begin{aligned}
&\mathbb{E}_{\widetilde{\mathcal{D}}_1,\widetilde{\mathcal{D}}_2}\left[(\widehat{\tau}^{[r]}(\widetilde{\mathbf{X}}_j \mid \widetilde{\mathcal{D}}_1) - \check{\tau}_j)^2\right] \\
&\quad = \underbrace{\mathbb{E}_{\widetilde{\mathcal{D}}_1,\widetilde{\mathcal{D}}_2}\left[(\widehat{\tau}^{[r]}(\widetilde{\mathbf{X}}_j \mid \widetilde{\mathcal{D}}_1) - \tau(\mathbf{X}_j))^2\right]}_{\text{True prediction error}} - 2\ \underbrace{\mathbb{E}_{\widetilde{\mathcal{D}}_1}\left[\widehat{\tau}^{[r]}(\widetilde{\mathbf{X}}_j \mid \widetilde{\mathcal{D}}_1) - \tau(\mathbf{X}_j)\right]}_{\text{Bias of } \widehat{\tau}^{[r]}(\widetilde{\mathbf{X}}_j)} \cdot \underbrace{\mathbb{E}_{\widetilde{\mathcal{D}}_2}\left[\check{\epsilon}_j\right]}_{\text{Bias of } \check{\tau}_j} + \underbrace{\mathbb{E}_{\widetilde{\mathcal{D}}_2}\left[\check{\epsilon}_j^2\right]}_{\text{Constant}}.
\end{aligned}$$

By Proposition 2.1, the bias of the AIPW score, $\mathbb{E}_{\widetilde{\mathcal{D}}_2}\left[\check{\epsilon}_j\right]$, is zero, which eliminates the second term. Since the third term remains constant throughout the post-processing procedure, the decomposition suggests that aligning the predictions from $\widetilde{\mathcal{D}}_1$ with the AIPW scores of individuals

in $\widetilde{\mathcal{D}}_2$ effectively reduces the true prediction error.

The key insight from this analysis is that each time post-processing is performed by minimizing the sum of squared residuals and assessing accuracy using the AIPW score, an independent dataset (i.e., an *honest* sample) must be used — one that was not involved in constructing the original model. This principle forms the foundation of the sample-splitting strategy embedded in our proposed post-processing algorithm.[6]

2.4.4 An Algorithm for Honest Post-processing

To incorporate the honesty principle into the post-processing procedure described in 2.4.2, we introduce two key modifications.

First, rather than using the entire dataset $\widetilde{\mathcal{D}}$ for training the initial CATE model, performing model adjustments, and determining stopping criteria, we split $\widetilde{\mathcal{D}}$ into three folds, ensuring that each task is performed on a separate fold. This prevents information leakage and mitigates the risk of overfitting.

Second, instead of using the same data to construct $\widehat{c}^{[r]}$ and determine the optimal step size $\rho^{[r]}$, we decouple these two steps by reserving separate subsets of data for each. This separation ensures that the step size is chosen independently of the specific adjustment model, preserving the honesty of the post-processing adjustments.

Sample Splitting.

We divide $\widetilde{\mathcal{D}}$ into three independent subsets: $\widetilde{\mathcal{D}}_{\text{train}}$, $\widetilde{\mathcal{D}}_{\text{adj}}$, and $\widetilde{\mathcal{D}}_{\text{val}}$. The first subset, $\widetilde{\mathcal{D}}_{\text{train}}$, is used exclusively to construct an initial CATE model. The second subset, $\widetilde{\mathcal{D}}_{\text{adj}}$, is reserved for adjusting the initial model based on the AIPW score. Finally, an additional subset, $\widetilde{\mathcal{D}}_{\text{val}}$, is reserved to determine whether to keep updating the current prediction model or terminate the algorithm. This validation set is essential because repeated post-processing can result in overfitting to the

[6]While conceptually similar, our approach differs fundamentally from existing sample-splitting methods in causal inference [155, 126, 180]. These methods use separate datasets to construct nuisance models (i.e., conditional mean outcome and propensity score models) to then estimate the CATE model. In contrast, our honest approach ensures that post-processing steps — such as adjustment model construction and step size selection — are performed on independent datasets. This prevents overfitting to the approximation errors of AIPW scores, thereby improving generalizability. See Appendix B.1.2 for a detailed literature review and comparison.

approximation error of individuals in $\widetilde{\mathcal{D}}_{\text{adj}}$, which compromises the model's accuracy. Using $\widetilde{\mathcal{D}}_{\text{val}}$ to guide updates or termination helps ensure that the model's performance remains robust.

Subgroup Cross-Learning

Next, we discuss how to ensure honesty when constructing the adjustment model and determining the step size in each iteration of the gradient boosting procedure. In standard gradient boosting methods, both steps are typically performed on the same dataset [e.g., 77, 57]. However, this can lead to overfitting, as $\rho^{[r]}$ may be unintentionally chosen to align the model's prediction error with the approximation error of the AIPW score (the second term in Equation 2.6).

Ideally, model adjustment and step-size determination should be performed on a separate dataset that has not been used for model construction or post-processing. However, doing this at each iteration would require an extremely large sample, which is often impractical. Therefore, we introduce a data-efficient strategy called *Subgroup Cross-Learning*. Specifically, we partition the adjustment set $\widetilde{\mathcal{D}}_{\text{adj}}$ into multiple subgroups before initiating the gradient boosting process. At each iteration, the adjustment model is trained on *one* subgroup, while the step size is determined by minimizing the squared residuals over the *remaining* subgroups. The CATE model is then updated using the subgroup adjustment model and step size that yield the greatest improvement in proxy squared residuals on a separate validation set. By ensuring that adjustment model fitting and step-size selection are based on statistically independent subsets, this approach effectively reduces the risk of overfitting.

The resulting procedure describes our proposed *Subgroup Cross-Learning* strategy:

(i) Before any post-processing, first split $\widetilde{\mathcal{D}}_{\text{adj}}$ into Q groups $(\widetilde{\mathcal{G}}^1_{\text{adj}}, \widetilde{\mathcal{G}}^2_{\text{adj}}, \ldots, \widetilde{\mathcal{G}}^Q_{\text{adj}})$.

(ii) For each iteration r, construct an adjustment model $\widehat{c}_q^{[r]}(\widetilde{\mathbf{X}}_j)$ separately for each subgroup $\widetilde{\mathcal{G}}^q_{\text{adj}}$.

(iii) Then, determine the step size $\rho_q^{[r]}$ for each subgroup model by minimizing the sum of squared residuals across the *rest* of the adjustment data:

$$\rho_q^{[r]} = \arg\min_{\rho} \sum_{j \in \widetilde{\mathcal{D}}_{\text{adj}} \setminus \widetilde{\mathcal{G}}^q_{\text{adj}}} \left[\widehat{\tau}^{[r-1]}(\widetilde{\mathbf{X}}_j) + \rho \widehat{c}_q^{[r]}(\widetilde{\mathbf{X}}_j) - \check{\tau}_j \right]^2, \tag{2.7}$$

For simplicity, we omit the notation $\widetilde{D}^{[r-1]}$ in $\widehat{\tau}^{[r-1]}(\cdot \mid \widetilde{D}^{[r]})$, where the $\widetilde{D}^{[r-1]}$ denotes the data used to construct $\widehat{\tau}^{[r-1]}$.

(iv) Identify the subgroup $\widetilde{\mathcal{G}}_{q^\star}$ that provides the greatest improvement in mean squared residuals for individuals in the validation set $\widetilde{\mathcal{D}}_{\text{val}}$. Specifically, select $q^\star$ as follows:

$$q^\star = \underset{q \in \{1,\cdots,Q\}}{\arg\min} \sum_{k \in \widetilde{\mathcal{D}}_{\text{val}}} \left[\widehat{\tau}^{[r-1]}(\widetilde{\mathbf{X}}_k) + \rho_q^{[r]} \widehat{c}_q^{[r]}(\widetilde{\mathbf{X}}_k) - \check{\tau}_k\right]^2.$$

(v) Update the CATE model using the selected subgroup calibration model and its corresponding step size: $\widehat{\tau}^{[r]}(\cdot) := \widehat{\tau}^{[r-1]}(\cdot) + \rho_{q^\star}^{[r]} \widehat{c}_{q^\star}^{[r]}(\cdot)$.

This approach ensures independence between the data used to train the adjustment model and the data used to determine the step size, effectively reducing the risk of overfitting due to approximation errors.

Finally, to appropriately determine the termination of the boosting process — i.e., stopping late enough to avoid underfitting yet early enough to prevent overfitting, we employ a commonly used early stopping method: the boosting process is terminated when no further improvement in predictive accuracy is observed on the validation set $\widetilde{\mathcal{D}}_{\text{val}}$. This serves as our early stopping criterion[7]:

$$\sum_{k \in \widetilde{\mathcal{D}}_{\text{val}}} \left[\widehat{\tau}^{[r-1]}(\widetilde{\mathbf{X}}_k) + \rho_{q^\star}^{[r]} \widehat{c}_{q^\star}^{[r]}(\widetilde{\mathbf{X}}_k) - \check{\tau}_i\right]^2 - \sum_{k \in \mathcal{D}_{\text{val}}} \left[\widehat{\tau}^{[r-1]}(\widetilde{\mathbf{X}}_k) - \check{\tau}_k\right]^2 > 0.$$

Summary.

Algorithm 1 presents the pseudo-code for the proposed method. By integrating the honesty principle through sample splitting (separating initial CATE prediction from post-processing) and subgroup cross-learning (separating adjustment model fitting from step-size determination), the algorithm reduces overfitting to approximation errors and enhances the accuracy of CATE predictions. To further enhance sample efficiency, one can apply cross-fitting [126]: swap the roles of $\widetilde{\mathcal{D}}_{\text{train}}, \widetilde{\mathcal{D}}_{\text{adj}}, \widetilde{\mathcal{D}}_{\text{val}}$, rerun Algorithm 1, and average the resulting predictions.

[7]Note that the threshold of improvement can also be set as a tuning parameter in the calibration algorithm. A larger threshold implies fewer updates being performed, acting as a form of regularization for the boosting process.

Algorithm 1 Subgroup Cross-Learning for Honest Model Calibration

Input: $Q \in \mathbb{N}$, $R \in \mathbb{N}$

Output: Post-processed model $\widehat{\tau}$

Data: LDP-protected Data $\widetilde{\mathcal{D}} = \{\widetilde{\mathcal{D}}_{\text{train}}, \widetilde{\mathcal{D}}_{\text{adj}}, \widetilde{\mathcal{D}}_{\text{val}}\}$; True Propensity Score $e(\widetilde{\mathbf{X}}_i)$

Construct the initial CATE model $\widehat{\tau}^{[0]}(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}_{\text{train}})$ using $\widetilde{\mathcal{D}}_{\text{train}}$ and generate predictions for $\widetilde{\mathcal{D}}_{\text{adj}}$.

Derive the AIPW score $\{\check{\tau}_j\}_{j \in \widetilde{\mathcal{D}}_{\text{adj}}}$ and $\{\check{\tau}_k\}_{k \in \widetilde{\mathcal{D}}_{\text{val}}}$ separately using the cross-fitting procedure.

Divide $\widetilde{\mathcal{D}}_{\text{adj}}$ into Q subgroups $(\widetilde{\mathcal{G}}^1_{\text{adj}}, \cdots, \widetilde{\mathcal{G}}^Q_{\text{adj}})$.

for $r = 1, \cdots, R$ **do**

Calculate the proxy residual error $\check{\tau}_j - \widehat{\tau}^{[r-1]}(\widetilde{\mathbf{X}}_j)$ for $j \in \widetilde{\mathcal{D}}_{\text{adj}}$.

for $q = 1, \cdots, Q$ **do**

(Adjustment Model Construction) Use individuals within the subgroup $\widetilde{\mathcal{G}}^q_{\text{adj}}$ to construct a machine learning model, denoted as $\widehat{c}^{[r]}_q(\widetilde{\mathbf{X}}_i)$, which is trained to predict $\check{\tau}_j - \widehat{\tau}^{[r-1]}(\widetilde{\mathbf{X}}_j)$ using $\widetilde{\mathbf{X}}_j$.

(Step Size Determination) Determine the step size by minimizing the squared proxy residuals across individuals in $\widetilde{\mathcal{D}}_{\text{adj}} \backslash \widetilde{\mathcal{G}}^q_{\text{adj}}$:

$$\rho^{[r]}_q = \arg\min_{\rho} \sum_{j \in \widetilde{\mathcal{D}}_{\text{adj}} \backslash \widetilde{\mathcal{G}}^q_{\text{adj}}} \left[\widehat{\tau}^{[r-1]}(\widetilde{\mathbf{X}}_j) + \rho \widehat{c}^{[r]}_q(\widetilde{\mathbf{X}}_j) - \check{\tau}_j \right]^2$$

end

Pick the subgroup that leads to the smallest squared proxy residuals in the validation set:

$$q^\star = \arg\min_{q \in \{1, \cdots, Q\}} \sum_{k \in \widetilde{\mathcal{D}}_{\text{val}}} \left[\widehat{\tau}^{[r-1]}(\widetilde{\mathbf{X}}_k) + \rho^{[r]}_q \widehat{c}^{[r]}_q(\widetilde{\mathbf{X}}_k) - \check{\tau}_k \right]^2.$$

Generate new predictions for the validation set: $\widehat{\tau}^{\text{new}}(\widetilde{\mathbf{X}}_k) = \widehat{\tau}^{[r-1]}(\widetilde{\mathbf{X}}_k) + \rho^{[r]}_{q^\star} \widehat{c}^{[r]}_{q^\star}(\widetilde{\mathbf{X}}_k),\ k \in \widetilde{\mathcal{D}}_{\text{val}}$.

if $\sum_{k \in \widetilde{\mathcal{D}}_{val}} \left[\widehat{\tau}^{new}(\widetilde{\mathbf{X}}_k) - \check{\tau}_k \right]^2 - \sum_{k \in \widetilde{\mathcal{D}}_{val}} \left[\widehat{\tau}^{[r-1]}(\widetilde{\mathbf{X}}_k) - \check{\tau}_k \right]^2 < 0$ **then**

Update the CATE model: $\widehat{\tau}^{[r]}(\cdot) = \widehat{\tau}^{[r-1]}(\cdot) + \rho^{[r]}_{q^\star} \widehat{c}^{[r]}_{q^\star}(\cdot)$.

end

else

Stop post-processing and return $\widehat{\tau}(\cdot) = \widehat{\tau}^{[r-1]}(\cdot)$.

end

end

2.4.5 Implementation Considerations

Our proposed solution allows for flexible model specifications and hyperparameter tuning. Below, we provide recommendations to guide both model selection and tuning.

Subgroup Partitioning Strategy.

In the proposed algorithm, subgroup cross-learning is essential for ensuring robustness and preventing overfitting. A key question is: What grouping strategies are most effective for improving predictive accuracy? Our findings suggest that when sample sizes are small, forming groups of homogeneous individuals is crucial for constructing reliable adjustment models.

There are multiple ways to achieve this objective. One effective approach is to divide individuals based on the quantiles of their initial CATE predictions ($\widehat{\tau}^{[0]}$). This strategy is motivated by theoretical insights from multicalibration boosting [32, 83], which suggest that minimizing error within groups of individuals with similar predicted values can improve accuracy.

Alternatively, subgroups can be formed using clustering algorithms to partition the covariate space based on individual characteristics or by grouping individuals according to their predicted outcome values. In Appendix B.5.5, we compare different subgrouping strategies and find that partitioning based on predicted CATEs yields the best performance when the experimental sample is small. However, with larger sample sizes, the choice of subgrouping strategy has minimal effect on overall performance.

Model Selection.

In the proposed algorithm, three model components need to be specified: the initial CATE model, $\widehat{\tau}^{[0]}$, the nuisance models in the AIPW score, $\widehat{\mu}_1$ and $\widehat{\mu}_0$, and the adjustment models in post-processing, $\left\{\widehat{c}^{[r]}\right\}$.

- **Initial CATE Model**: As we show through simulations and real-world data, our proposed solution improves predictive accuracy and treatment prioritization, regardless of the initial CATE model. In practice, we recommend selecting the initial CATE model that achieves the highest predictive accuracy or treatment prioritization performance before calibration, as our analyses consistently show that such a model tends to retain its superior performance even after

calibration. This initial selection can be efficiently implemented using the bootstrap validation procedure described in Section 2.6.4.

- **Nuisance Models in AIPW score**: The AIPW score construction involves two nuisance models, $\widehat{\mu}_1$ and $\widehat{\mu}_0$, for the conditional mean outcomes. Following [155], we recommend selecting these models based on the highest predictive accuracy.

- **Select the Class of Adjustment Models**: In theory, any model class can be used to construct the adjustment models. However, based on our experiments with both simulated and real-world data, we find that simpler models (such as linear regression) often outperform more complex ones (e.g., regression forests) in terms of both training speed and accuracy. Thus, we recommend using simpler models to fit the adjustment models. This recommendation aligns with the philosophy of boosting [77], which aims to improve predictive performance by combining the outputs of multiple weak models.

Hyperparameter Tuning

Our algorithm has two primary hyperparameters: the number of subgroups (Q) and the number of iterations (R).

- *Number of Subgroups* (Q): The choice for the number of subgroups presents an inherent bias-variance trade-off. If Q is too small, the algorithm may not effectively identify those individuals most informative for model correction, resulting in a larger bias. Conversely, if Q is too large, the algorithm can become unstable, increasing the variance.[8] Overall, we recommend that each subgroup should consist of at least 100 individuals/observations to ensure algorithm stability.

- *Number of Iterations* (R): Through simulation analyses, we have found that setting $R = Q$ is typically sufficient. The reason behind this recommendation is that the step size for subgroup q often approaches zero when it has been previously used for updating the initial model. In other words, once the algorithm has utilized a specific subgroup to update the predictions, the value of updating again based on that subgroup diminishes considerably.

[8]This trade-off is further demonstrated using simulation in Appendix B.5.1.

2.5 Empirical Performance: Simulation

We conduct simulation analyses for two main purposes. First, to compare the performance of our proposed method against several benchmarks under LDP protection. Second, to evaluate how our algorithm and competing methods perform across varying data sizes, providing insight into their large-sample behavior.

2.5.1 Simulation Setup

We generate an experimental sample with a binary treatment variable ($W_i \in \{0, 1\}$) where the treatment assignment is completely random, with equal proportions for both treatment and control groups. The outcome variable is generated according to the following process:

$$
\begin{aligned}
Y_i(W_i) &= b(\mathbf{X}_i) + (W_i - 0.5)\tau(\mathbf{X}_i) + \epsilon_i,\ \epsilon_i \sim_{i.i.d.} \mathcal{N}(0, 5), \\
b(\mathbf{X}_i) &= \frac{1}{4}\left[\sin(\pi X_{i,1} X_{i,2}) - 2(X_{i,1} - X_{i,3} - 0.5)^2 + X_{i,2} X_{i,4} + 2X_{i,5}^2 + X_{i,6}^2\right], \\
\tau(\mathbf{X}_i) &= \frac{1}{4}\left[2X_{i,1} - X_{i,3}\cos(\pi X_{i,2}) + 0.5X_{i,3}^2 - X_{i,4} - \log(X_{i,1})(X_{i,4} - 2.5) - 8\right],
\end{aligned}
$$

where each covariate is identically and independently distributed (i.i.d.) from Uniform$(0, 5)$.

We analyze two scenarios under the data-generating process described above.[9] In the first, covariates are protected by LDP, meaning only their noisy versions are observed: $\widetilde{X}_{i,p} = X_{i,p} + \eta_{i,p}$, where $\eta_{i,p} \sim$ Laplace$(0, \sigma)$ are i.i.d Laplace noises. Additionally, for each bootstrap sample, fresh noise is injected into the true covariates of the holdout individuals to generate CATE predictions. The second scenario assumes LDP protection for the outcome variable, so only the noisy outcome is observed: $\widetilde{Y}_i(W_i) = Y_i(W_i) + \eta_i$, where $\eta_i \sim$ Laplace$(0, \sigma)$. For both scenarios, we vary σ to examine the privacy-accuracy trade-off and adjust the experimental sample size to evaluate sample efficiency.

2.5.2 Methods for Comparison

We compare the performance of our proposed solution against three alternatives, including a baseline using a standard CATE model and two simplified versions of our approach. This com-

[9] In Appendix B.5.4, we present results for the setting where both covariates and outcomes are protected by LDP. The findings are consistent with the main results reported in the paper.

parison not only establishes the superiority of our method but also quantifies the improvements gained from its individual components. We evaluate a total of four estimation methods:

1. **Default** (DEFAULT): This method constructs a CATE model using the experimental dataset $\widetilde{\mathcal{D}} = \{\widetilde{\mathcal{D}}_{\text{train}}, \widetilde{\mathcal{D}}_{\text{adj}}, \widetilde{\mathcal{D}}_{\text{val}}\}$ without any sample splitting. It represents the standard approach typically adopted by decision-makers for CATE estimation without any additional post-processing.

2. **Non-honest Post-Processing** (NON-HONEST): This method applies post-processing by performing traditional boosting on the initial CATE model using the full dataset $\widetilde{\mathcal{D}}$, without any sample splitting or subgroup cross-learning. Essentially, it is equivalent to existing CATE estimation frameworks from [155] and [126], with boosting methods as the model class for CATE modeling.

3. **Post-Processing with Sample Splitting** (SPLIT-ONLY): This method performs post-processing using a separate adjustment dataset $\widetilde{\mathcal{D}}_{\text{adj}}$ and stops when there is no improvement on a third validation set $\widetilde{\mathcal{D}}_{\text{val}}$. However, it determines the adjustment model and step size using all individuals in $\widetilde{\mathcal{D}}_{\text{adj}}$ without subgroup cross-learning.

4. **Post-processing with Sample Splitting and Subgroup Cross-Learning** (PROPOSED): This method implements the proposed honest subgroup cross-learning approach for post-processing, in which subgroups are formed based on the predicted values from the initial CATE model.

2.5.3 Estimation Details

This section outlines the key details of our model implementation. Additional specifications and technical details are provided in Appendix B.5.1.

In our main analysis, we use the R-learner with regression forests [155] as the DEFAULT method as well as the initial CATE model across all approaches. The AIPW score is computed using OLS regression with squared and interaction terms for the conditional mean outcome models, as this specification provides the best predictive accuracy among the candidate models tested. Similarly, the adjustment models used in post-processing are also estimated via OLS regression with squared and interaction terms.

For the PROPOSED and SPLIT-ONLY methods, the experimental dataset is divided into three equal-sized folds, ensuring that cross-fitting is applied for robustness.

2.5.4 Evaluation Procedure

We evaluate model performance using $B = 100$ bootstrap replications and report the mean of key metrics. First, we generate a holdout set $\widetilde{\mathcal{D}}_{\text{holdout}}$, consisting of $N_{\text{holdout}} = 10,000$ individuals, which is not involved in any model estimation or post-processing but only used for model evaluation across all bootstrap replications. In each replication b, we generate an experimental set $\widetilde{\mathcal{D}}^b$ and use it to estimate CATE using each of the four different methods. Predictions are then generated on the holdout set, and prediction errors are calculated for each method. Specifically, the prediction error for individual $l \in \widetilde{\mathcal{D}}_{\text{holdout}}$ is defined as $\text{err}_l^b = \widehat{\tau}_l^b - \tau(\mathbf{X}_l)$, where $\widehat{\tau}_l^b$ denotes the final prediction of individual l in the b-th replication. After completing this process across $B = 100$ bootstrap replications, we compute the mean squared error (MSE)[10] for each method as follows:

$$\widehat{\mathbb{E}}\left[(\text{err}_l^b)^2\right] = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{N_{\text{holdout}}}\sum_{l\in\widetilde{\mathcal{D}}_{\text{holdout}}}\left[\text{err}_l^b\right]^2.$$

Beyond evaluating statistical accuracy, we assess the improvement in treatment prioritization performance achieved by our proposed method when decision-makers rely on CATE models estimated from LDP-protected data. Specifically, we assess this improvement using the Area Under the Targeting Operating Characteristic curve (AUTOC) [211], a popular metric that gauges a model's ability to correctly rank individuals based on their treatment effects. Specifically, given the predicted CATEs $\widehat{\tau}(\mathbf{X}_i)$ for i in the holdout set, the *Targeting Operating Characteristic* (TOC) is defined as the difference in treatment effects between individuals within the top $\phi \times 100\%$ predicted CATE tier and all individuals, i.e.,

$$\text{TOC}(\phi;\widehat{\tau}) = \mathbb{E}\left[Y_i(1) - Y_i(0) | F_{\widehat{\tau}}(\widehat{\tau}_i) \geqslant 1-\phi\right] - \mathbb{E}\left[Y_i(1) - Y_i(0)\right], \tag{2.8}$$

where $F_{\widehat{\tau}}$ is the cumulative distribution function of the predicted CATEs, and $\widehat{\tau}_i$ is defined as $\widehat{\tau}(\widetilde{\mathbf{X}}_i)$ when the covariates are protected by LDP and as $\widehat{\tau}(\mathbf{X}_i)$ when the outcome is protected by LDP.

[10] In Appendix B.5.2, we also present the mean squared bias and variance for model comparison.

Then, the AUTOC is defined as

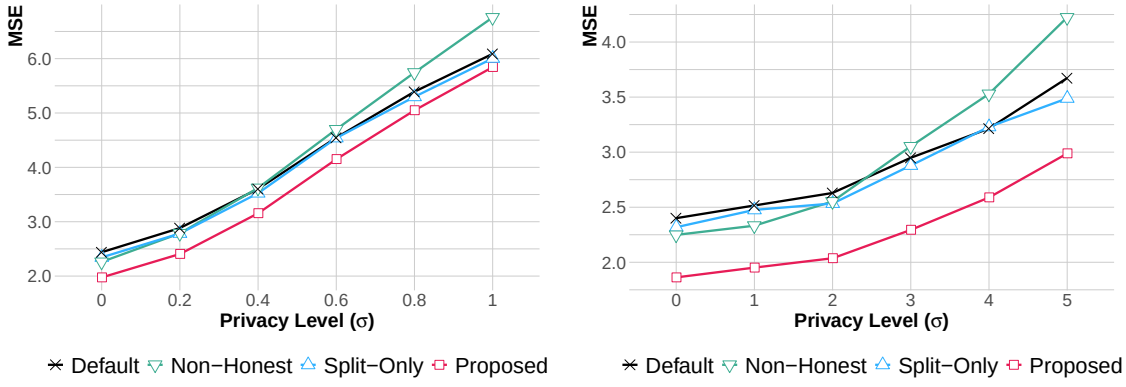
$$\text{AUTOC}(\widehat{\tau}) = \int_0^1 \text{TOC}(\phi; \widehat{\tau}) d\phi. \tag{2.9}$$

Note that a model $\widehat{\tau}$ outperforms another model $\widehat{\tau}'$ in identifying individuals in the top $\phi \times 100\%$ CATE group if $\text{TOC}(\phi; \widehat{\tau}) > \text{TOC}(\phi; \widehat{\tau}')$. Therefore, a higher AUTOC value indicates that the CATE model more effectively identifies individuals with the greatest incremental impact from the intervention and produces better treatment prioritization rules.[11]

2.5.5 Results

Varying Privacy Level.

Figure 2.3 illustrates the MSE of different methods across a range of privacy levels ($\sigma \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ for the covariates and $\sigma \in \{1, 2, 3, 4, 5\}$ for the outcome variable). Specifically, Figure 2.3a presents the results when the covariates are protected by LDP, and Figure 2.3b depicts the results when the outcome is protected by LDP. These results are generated using a sample of 3,000 individuals in the experimental set to estimate CATE models, and their performances are evaluated using a holdout set of 10,000 individuals.



(a) *Scenario 1: LDP-Protected Covariates* **(b)** *Scenario 2: LDP-Protected Outcome*

Note: We simulate 100 replications to compute the bootstrap mean MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point. The results presented here are based on using R-learner with regression forests as the initial CATE model. Results derived from different CATE models are available in Appendix B.5.6.

Figure 2.3: *Mean Squared Error of Different Methods Across Varying Privacy Levels*

[11] In Appendix B.5.2, we evaluate the value improvement of the PROPOSED method for a simple targeting rule that assigns treatment to individuals with positive predicted CATEs.

We identify several key patterns in the results. First, there exists reveal clear trade-offs between privacy and accuracy: as the privacy level increases, the prediction error also increases, regardless of the method employed. Second, the Proposed method consistently achieves lower prediction errors across all privacy levels, regardless of whether LDP protection is applied to covariates or outcomes. Third, the Split-only and Non-honest approaches do not improve the accuracy of prediction, with their MSE even exceeding that of the Default approach. Interestingly, while the Split-only approach performs similarly to the Non-Honest method at lower privacy levels (due to the smaller sample size used for the construction of $\widehat{\tau}^{[0]}$), it outperforms the Non-Honest approach at higher privacy levels. This underscores the risk of overfitting in post-processing without honesty, particularly under stricter privacy protection settings. Together, these findings highlight (i) the importance of sample splitting and (ii) the value of the proposed subgroup cross-learning method for effective model calibration.

Next, we assess the effectiveness of treatment prioritization rules derived from predicted CATEs across different methods. Table 2.2 presents the AUTOC values, along with the proportion of bootstrap replications (in parentheses) where each method outperforms the Default approach across different privacy levels. The results suggests four key findings. First, consistent with previous results, there is a clear privacy-utility trade-off, as AUTOC values decline with increasing privacy protection. Second, the Proposed method consistently outperforms all alternatives at all privacy levels, demonstrating its superior ability to optimize treatment prioritization under LDP protection. Third, the Split-only method improves AUTOC values compared to the Default approach. This result suggests that if a decision-maker adopts a relative targeting rule — such as treating individuals with the highest predicted CATEs — the Split-only method enhances targeting performance relative to the Default method. Finally, the Non-Honest method fails to demonstrate any improvement in treatment prioritization. Collectively, these findings highlight the importance of maintaining honesty in post-processing CATE predictions and demonstrate the effectiveness of the Proposed method in enhancing treatment allocation.

Table 2.2: *AUTOC Values for Different Methods Across Varying Privacy Levels*

(a) Scenario 1: LDP-Protected Covariates

σ	Proposed	Split-only	Non-Honest	Default
0.0	2.39 (100%)	2.36 (93%)	2.34 (68%)	2.32
0.2	2.31 (100%)	2.27 (84%)	2.25 (75%)	2.23
0.4	2.17 (100%)	2.14 (90%)	2.11 (76%)	2.08
0.6	1.99 (100%)	1.96 (94%)	1.90 (53%)	1.90
0.8	1.80 (100%)	1.77 (94%)	1.71 (58%)	1.70
1.0	1.61 (96%)	1.59 (92%)	1.50 (31%)	1.52

(b) Scenario 2: LDP-Protected Outcome

σ	Proposed	Split-only	Non-Honest	Default
0	2.42 (96%)	2.39 (84%)	2.38 (71%)	2.36
1	2.42 (99%)	2.37 (83%)	2.36 (76%)	2.34
2	2.40 (97%)	2.37 (84%)	2.34 (63%)	2.32
3	2.37 (98%)	2.32 (88%)	2.28 (68%)	2.26
4	2.32 (96%)	2.29 (92%)	2.23 (61%)	2.22
5	2.25 (98%)	2.23 (95%)	2.14 (68%)	2.13

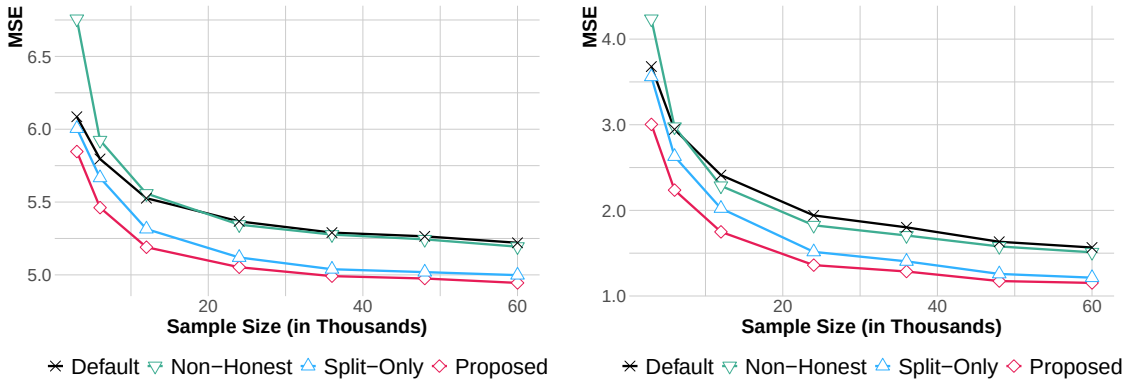
Note: We calculate the AUTOC values from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses). The results presented here are based on using R-learner with regression forests as the initial CATE model. Results derived from different CATE models are available in Electronic Companion B.5.6.

Varying Sample Sizes.

Next, we compare different methods across varying sizes of experimental data (ranging from 3,000 to 60,000 individuals) while keeping the privacy levels fixed ($\sigma = 1.0$ for the covariates and $\sigma = 5$ for the outcome variable). Figure 2.4 presents the mean squared error of the different methods estimated across varying sample sizes. The results highlight several key findings.

First, as expected, the MSE decreases as the sample size increases. Notably, our Proposed method (represented by the red line) consistently achieves the lowest MSE across all sample sizes. These results demonstrate the value of our proposed methods, even when the company is able to collect a large number of samples. Secondly, when the sample size is sufficiently large, both the Split-only (blue triangle line) and Non-Honest (green line) approaches outperform the Default approach (black line). This highlights the benefits of model calibration, particularly when there is sufficient data to effectively inform and reduce errors. Third, the Split-only approach significantly outperforms the Non-Honest method when the sample size is sufficiently large. Besides, as the sample size increases, the performance of Split-only converges closer to that of Proposed.

Together, these results demonstrate the importance of incorporating honesty in the post-processing procedure. First, in large samples, splitting data between initial CATE modeling and post-processing is crucial for preventing overfitting. Second, in smaller datasets, subgroup cross-learning plays a particularly valuable role in improving model performance by enhancing stability and reducing variance.



(a) *Scenario 1: LDP-Protected Covariates* **(b)** *Scenario 2: LDP-Protected Outcome*

Note: We simulate 100 replications to compute the bootstrap mean MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point. The results presented here are based on using R-learner with regression forests as the initial CATE model. Results derived from different CATE models are available in Appendix B.5.6.

Figure 2.4: *Mean Squared Error of Different Methods with Varying Sample Size*

Next, Table 2.3 presents the AUTOC values for different methods across varying sample sizes. Consistent with previous findings, the Proposed method consistently outperforms all alternatives in treatment prioritization, achieving the highest AUTOC values across all samples. Second, when sample sizes are large, both the Split-only and Non-Honest methods can outperform the Default approach, with Split-only achieving performance which is comparable to the Proposed specification. Together, these results highlight the benefits of well-designed sample splitting and post-processing when data is protected by LDP, demonstrating their effectiveness compared to the Default method even when the decision-makers have access to large experimental samples.

Table 2.3: *AUTOC Values for Different Methods Across Varying Sample Sizes*

(a) Scenario 1: LDP-Protected Covariates

Sample Size	Proposed	Split-only	Non-Honest	Default
3k	1.61 (96%)	1.59 (92%)	1.50 (32%)	1.52
6k	1.67 (100%)	1.66 (100%)	1.59 (70%)	1.58
12k	1.72 (100%)	1.71 (100%)	1.66 (90%)	1.65
24k	1.75 (100%)	1.74 (100%)	1.70 (95%)	1.68
36k	1.76 (100%)	1.75 (100%)	1.71 (100%)	1.70
48k	1.76 (100%)	1.76 (100%)	1.72 (100%)	1.70
60k	1.77 (100%)	1.76 (100%)	1.72 (100%)	1.71

(b) Scenario 2: LDP-Protected Outcome

Sample Size	Proposed	Split-only	Non-Honest	Default
3k	2.25 (98%)	2.23 (96%)	2.14 (68%)	2.13
6k	2.37 (100%)	2.34 (98%)	2.29 (92%)	2.25
12k	2.44 (100%)	2.42 (98%)	2.37 (90%)	2.34
24k	2.49 (100%)	2.48 (100%)	2.43 (100%)	2.40
36k	2.50 (100%)	2.49 (100%)	2.45 (100%)	2.43
48k	2.52 (100%)	2.51 (100%)	2.47 (100%)	2.45
60k	2.52 (100%)	2.51 (100%)	2.48 (100%)	2.46

Note: We calculate the AUTOC values from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses). The results presented here are based on using R-learner with regression forests as the initial CATE model. Results derived from different CATE models are available in Appendix B.5.6.

2.5.6 Robustness Checks

In Appendix B.5.6, we conduct robustness checks using alternative model classes for the Default approach and initial CATE models. We find that the Proposed approach consistently outperforms regardless of the chosen model class. Additionally, we present results from other post-processing methods, including various subgroup partitioning strategies (Appendix B.5.5) for the Proposed method and alternative post-processing techniques (Appendix B.5.5). Overall, our proposed method—using subgroups based on initial CATE predictions—consistently delivers the best performance, with covariate-based subgrouping also yielding substantial gains.

2.6 Empirical Performance: Real-world Case Studies

We validate the proposed solution using two real-world applications that serve as established benchmarks for CATE models [e.g., 170]. Both studies involve randomized marketing campaigns designed to encourage customer purchases.

2.6.1 Studies Overview

Study 1: Hillstrom E-mail Campaign.

The first study uses the Hillstrom dataset, originally from the MineThatData E-Mail Analytics and Data Mining Challenge [102]. This dataset includes 64,000 customers, who were randomly assigned to one of three groups in an e-mail marketing experiment: a group that received an e-mail promoting men's merchandise, a group that received an e-mail promoting women's merchandise, and a control group that received no e-mail campaign. Following prior research using this dataset [123, 54, 170], we focus on evaluating the effectiveness of the women's merchandise e-mail promotion compared to no e-mail at all. Our final sample consists of 42,693 customers, evenly split between the treatment group (21,387 customers) and the control group (21,306 customers). Further data descriptions can be found in Appendix B.6.

Study 2: Starbucks Promotional Campaign Data.

The second study uses data from a promotional campaign conducted through the Starbucks Rewards mobile app, made available by the Udacity Data Science Program. This dataset captures an experiment in which a subset of customers was randomly offered a promotion (the intervention) to encourage product purchases (the outcome variable). The dataset includes 126,184 customers with seven anonymous pre-treatment covariates. The sample is evenly split, with 63,112 customers in the treatment group and 63,072 in the control group. The response rates are notably low — 1.68% in the treatment group and 0.73% in the control group, resulting in an average treatment effect of 0.95%. Additional details can be found in Appendix B.7.

2.6.2 Implementation of LDP

Since these datasets are not originally protected by LDP, we simulate two scenarios implementing the most common LDP methods: one where pre-treatment covariates are protected by LDP and another where the outcome is protected by LDP. For discrete variables, we apply the randomized response mechanism[62]. Under this mechanism, the true value is observed with probability $1 - f$, while with probability f, a random draw from all possible values (each appearing with equal probability) is observed instead. For continuous variables, we introduce the Laplace(σ) noise. The privacy level is varied from a "Very Low" scenario (i.e., small f and σ) to a "Very High" (i.e., large f and σ) scenario. Additional details about the implementation can be found in Appendix B.6.3 and B.7.3.

2.6.3 Methods for Comparison and Estimation Details

As in Section 2.5, we compare the Proposed method with the Default method, which estimates the CATE model directly without post-processing, as well as the Non-honest and Split-only methods. The model specifications closely follow those described in Section 2.5, with the following key details.

In the main analysis, we use the R-learner with regression forests [155] as the Default method and as the initial CATE model across different approaches. The AIPW score is computed using OLS regression for the conditional mean outcome models, as this specification provides the highest

predictive accuracy among the candidate models tested. Similarly, the adjustment models $\widehat{c}_q^{[r]}$ used in post-processing are also constructed via OLS regression. For detailed model specifications and the complete set of results across different DEFAULT models, please refer to Appendix B.6 and B.7.

2.6.4 Performance Evaluation

To evaluate the performance of each method, we employ a bootstrap validation scheme similar to that used by **(author?)** [9]. Specifically, we generate $B = 50$ random splits, each consisting of a model construction set (70%) and a holdout set (30%). For each split, we estimate CATE models using the four methods (PROPOSED, DEFAULT, NON-HONEST and SPLIT-ONLY) using the model construction set and predict CATEs for all individuals in the holdout set. Note that when covariates are protected by LDP, we inject noise into both the experimental and holdout sets to fully capture the impact of noise on model training and predictions. However, when the outcome variable is protected by LDP, we do not inject noise into the holdout set outcomes. This ensures that the performance metric reflects the true value improvement while reducing variance in the evaluation.

Unlike in our simulation analyses, where the true CATE for each individual is directly observable, real-world data lacks this ground truth. Therefore, we assess the predictive accuracy of treatment effects by comparing the group average treatment effect (GATE) between the actual observed outcomes and the model predictions. Specifically, we first divide the holdout set into ten equally sized groups based on the predicted CATE rankings from a given model for a random split. Next, we compute the GATE based on (i) the predicted CATEs and (ii) the actual outcomes observed in the experiment. We then compute the root mean squared error (RMSE) between these two metrics across the ten groups and average the results over 50 random splits. Additionally, we evaluate the treatment prioritization ability of different methods by computing the AUTOC value on the holdout set.

2.6.5 Results

Table 2.4 presents the RMSE of GATE (multiplied by 100) for different methods, along with the percentage of random splits where the focal method achieves a lower RMSE compared to

the DEFAULT method. The results are reported across a range of privacy levels, from "Very Low" to "Very High", under two scenarios: differentially-private covariates and differentially-private outcomes.

Table 2.4: *RMSE of GATE (Multiplied by 100) for Different Methods Across Varying Privacy Levels*

(a) Study 1: Hillstrom Data

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
No	3.96 (100%)	3.98 (100%)	7.29 (52%)	7.29	3.96 (100%)	3.98 (100%)	7.29 (52%)	7.29
Very Low	3.14 (100%)	3.19 (100%)	5.20 (52%)	5.20	4.35 (100%)	4.43 (100%)	7.88 (44%)	7.86
Low	3.39 (100%)	3.43 (100%)	5.56 (46%)	5.55	4.39 (100%)	4.43 (100%)	8.18 (36%)	8.14
Medium	3.47 (100%)	3.55 (100%)	5.61 (60%)	5.62	4.92 (100%)	5.05 (100%)	8.76 (40%)	8.71
High	3.47 (100%)	3.50 (100%)	5.61 (34%)	5.56	4.99 (100%)	5.04 (100%)	8.91 (40%)	8.90
Very High	3.44 (100%)	3.48 (100%)	5.62 (42%)	5.60	5.10 (100%)	5.16 (100%)	9.33 (40%)	9.29

(b) Study 2 : Starbucks Data

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
No	0.71 (100%)	0.71 (100%)	1.40 (32%)	1.38	0.71 (100%)	0.71 (100%)	1.40 (32%)	1.38
Very Low	0.73 (100%)	0.74 (100%)	1.42 (33%)	1.41	1.59 (100%)	1.60 (100%)	2.85 (36%)	2.85
Low	0.78 (100%)	0.78 (100%)	1.44 (42%)	1.43	2.12 (100%)	2.14 (100%)	3.71 (48%)	3.71
Medium	0.80 (100%)	0.82 (100%)	1.47 (44%)	1.47	2.52 (100%)	2.52 (100%)	4.36 (48%)	4.36
High	0.84 (100%)	0.86 (100%)	1.47 (33%)	1.45	2.82 (100%)	2.84 (100%)	4.85 (36%)	4.84
Very High	0.84 (100%)	0.84 (100%)	1.42 (52%)	1.42	3.06 (100%)	3.09 (100%)	5.27 (32%)	5.27

Note: We calculate the RMSE from 50 random splits, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses). The results presented here are based on the use of R-learner with regression forests as the initial CATE model. Results derived from different CATE models are available in Appendices B.6.7 and B.7.

First, while RMSE increases with stricter privacy protection when the outcome of interest is protected by LDP, this trend does not necessarily hold when covariates are protected. At first glance, this may seem surprising, but it aligns with our bias-variance characterization in Section 2.3 — when covariate noise is present, the CATE model detects less heterogeneity, resulting in more conservative CATE predictions. In fact, in the Hillstrom dataset, we observe that the variance of predicted CATEs from the DEFAULT method shrinks by 20% under "Very Low" covariate privacy protection compared to the no-privacy-protection case. Second, consistent with the theoretical insights from Section 2.4 and the simulation results in Section 2.5, both the PROPOSED and SPLIT-ONLY methods improve predictive accuracy compared to the DEFAULT method, while the NON-HONEST method fails to do so.

Table 2.5 presents the AUTOC values (multiplied by 100) for different methods across a range of privacy levels, under two scenarios: differentially-private covariates and differentially-private outcomes. First, as privacy protection increases, AUTOC values decline, highlighting the privacy-

utility trade-off in resource allocation for targeted interventions. Second, all post-processing methods consistently outperform the Default method, reinforcing post-processing as a promising framework for enhancing treatment prioritization when data is LDP-protected. Third, consistent with previous results, both the Proposed and Split-only methods outperform the Non-Honest methods, with the Proposed method performing slightly better than the Split-only method. These results provide empirical support for the importance of honesty in post-processing CATE predictions.

Table 2.5: *AUTOC Values (Multiplied by 100) for Different Methods Across Varying Privacy Levels*

(a) Study 1: Hillstrom Data

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Split-only	Non-Honest	Default	Proposed	Split-only	Non-Honest	Default
No	1.38 (82%)	1.36 (82%)	1.18 (60%)	1.17	1.37 (86%)	1.38 (82%)	1.18 (60%)	1.17
Very Low	1.46 (92%)	1.42 (88%)	1.17 (80%)	1.13	1.10 (76%)	1.07 (72%)	0.92 (54%)	0.91
Low	1.24 (94%)	1.23 (98%)	0.98 (82%)	0.92	1.11 (82%)	1.08 (78%)	0.90 (70%)	0.87
Medium	1.12 (90%)	1.09 (86%)	0.83 (84%)	0.76	0.99 (74%)	0.95 (68%)	0.84 (62%)	0.80
High	0.98 (92%)	0.95 (90%)	0.74 (84%)	0.69	0.97 (78%)	0.96 (66%)	0.85 (72%)	0.81
Very High	0.94 (90%)	0.91 (88%)	0.69 (84%)	0.62	0.75 (86%)	0.73 (82%)	0.56 (62%)	0.52

(b) Study 2 : Starbucks Data

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Split-only	Non-Honest	Default	Proposed	Split-only	Non-Honest	Default
No	0.45 (96%)	0.44 (96%)	0.38 (64%)	0.38	0.45 (96%)	0.44 (96%)	0.38 (64%)	0.38
Very Low	0.43 (94%)	0.42 (86%)	0.37 (84%)	0.35	0.20 (84%)	0.20 (84%)	0.16 (92%)	0.14
Low	0.33 (92%)	0.32 (78%)	0.29 (84%)	0.27	0.13 (84%)	0.13 (80%)	0.11 (100%)	0.08
Medium	0.25 (92%)	0.24 (92%)	0.21 (96%)	0.19	0.11 (88%)	0.08 (76%)	0.06 (100%)	0.04
High	0.19 (78%)	0.19 (76%)	0.18 (78%)	0.16	0.09 (76%)	0.09 (68%)	0.07 (92%)	0.05
Very High	0.19 (74%)	0.18 (74%)	0.17 (76%)	0.15	0.09 (80%)	0.09 (68%)	0.08 (92%)	0.06

Note: We calculate the AUTOC values from 50 random splits, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses). The results presented here are based on the use of R-learner with regression forests as the initial CATE model. Results derived from different CATE models are available in Appendixs B.6.7 and B.7.

While the performance of different post-processing methods is relatively comparable in these scenarios — likely due to the large sample sizes in both datasets — we also evaluate a small-sample setting, using only 10% of the data for model construction and 90% for holdout evaluation (see Appendix B.6.5 and B.7.5). Consistent with our simulation findings, when sample size is small, the Proposed method significantly outperforms all other methods in both predictive accuracy and treatment prioritization. The Split-Only method also improves upon the Default, while the Non-Honest method tends to hurt predictive accuracy (though it still offers some gains in treatment prioritization).

2.6.6 Robustness Checks

In Appendixs B.6.7 and B.7, we present robustness checks using alternative model classes for the Default approach and initial CATE models. The Proposed method consistently outperforms the Default in both predictive accuracy and treatment prioritization, and generally outperforms the Split-Only and Non-Honest methods, especially in noisy or data-constrained settings. While the Split-Only method offers some gains, the Non-Honest method often provides little to no improvement. The only minor exception is in the Starbucks dataset using the T-learner with XGBoost models, where all methods yield similarly high AUTOC values (compared to other CATE models).[12] Nevertheless, the Proposed method still achieves a substantial reduction in RMSE, highlighting its effectiveness in refining estimates even when the initial model ranks well.

In sum, the robustness checks highlight the importance of honesty in post-processing and demonstrate the consistent performance improvement of the Proposed method across a variety of model classes.

2.7 Discussion and Future Directions

Local differential privacy has gained significant attention and adoption in recent years as a robust framework for protecting individual data while enabling large-scale data collection and personalization. However, its impact on the precision of existing CATE models and firms' treatment prioritization capabilities remains largely unexplored. This paper takes the first step in investigating this issue and demonstrates — both theoretically and empirically — that the noise introduced by LDP can substantially degrade the predictive accuracy of CATE models and hinder effective resource prioritization. These challenges pose significant hurdles for organizations that rely on privacy-protected data for personalized interventions.

While existing methods for correcting measurement error bias exist, they are not well-suited for CATE prediction under LDP protection. Many rely on strict modeling assumptions that limit their applicability, while others require additional clean data, contradicting the privacy protection objective. To address these challenges, we propose a novel post-processing approach that

[12]This is likely because the best cross-validated XGBoost models in this setting are already simple and less prone to high variance, leaving limited room for improvement in treatment prioritization.

systematically refines CATE predictions using an unbiased CATE proxy, without compromising privacy guarantees. Importantly, by incorporating the *honesty* principle into the algorithm, we mitigate overfitting to the proxy's approximation errors, which can be substantial under LDP protection.

Our work has several limitations, which open up promising directions for future research. First, while we have characterized the impact of LDP on CATE prediction and treatment prioritization, we cannot directly observe the actual profitability loss due to constraints in our empirical data. A systematic analysis of privacy costs — such as evaluating multiple companies' performance metrics before and after LDP implementation — would provide more insights for policy decisions. In addition, our study focuses narrowly on CATE prediction, whereas LDP's impacts extend across various marketing applications. Future research could explore its implications for other marketing problems, such as demand estimation and product recommendations.

Second, our focus on treatment prioritization rules derived from predicted CATEs represents only one approach to targeting. The impact of LDP on alternative targeting methodologies and potential solutions requires further investigation. For instance, policy learning frameworks [e.g., 133, 17] aim to learn treatment rules that optimally classify individuals based on whether their treatment effect exceeds a specific decision threshold. In such contexts, minimizing mean-squared prediction error becomes secondary to enhancing classification accuracy. A promising research direction would be to modify the learning objectives within boosting procedures and develop adjustment models using alternative loss functions — such as cross-entropy loss [144] or expected profit loss [39] — that align more directly with policy learning objectives. Exploring these modifications would yield valuable insights into designing robust targeting policies that maintain effectiveness under LDP constraints while optimizing for the specific decision-making objectives of firms.

Third, while our solution is motivated by the need for a model-agnostic and privacy-preserving approach to correcting heterogeneous prediction errors under LDP protection, it extends more broadly as a general framework for CATE estimation in settings where experimental data is noisy or subject to measurement errors. Our simulations and empirical applications indicate that the proposed method enhances predictive accuracy and improves treatment prioritization, even in *the absence of LDP noise injection*– likely due to its ability to mitigate inherent noise in the data.

However, in non-LDP settings, existing measurement error correction methodologies may also be adaptable for improving accuracy, as they do not need to account for privacy constraints. A comprehensive comparison between these approaches and our proposed solution is beyond the scope of this research but presents an interesting avenue for future research.

Fourth, our approach requires learning from an unbiased AIPW score of the true CATE, which can be obtained when the decision-maker conducts a fully randomized experiment or assigns treatments based on a known propensity score. However, when the true propensity score is unknown (e.g., in advertising campaigns where third-party platforms or automated bidding systems determine ad recipients instead of the decision-maker, its estimates may become biased due to LDP noise in the covariates. In such cases, the AIPW estimator is no longer an unbiased signal for the true CATE, and an additional bias correction step for the propensity scores is needed to ensure the validity of our post-processing framework. Investigating effective bias correction methods for propensity scores in the presence of LDP-induced noise is an important direction for future research.

Finally, while we have demonstrated that the proposed method enhances CATE predictions through theoretical arguments, Monte Carlo simulations, and empirical examples, an important open question remains: whether and under what conditions the true CATE function is identifiable under LDP protection. [177] establishes formal identification results for classical error-in-variable models with a single covariate, except for certain parametric functions. [24] highlights that covariate measurement error can introduce bias in treatment effect estimation, particularly under non-random treatment assignment. Future research could further explore identification conditions of CATE under LDP constraints, develop new correction techniques, and assess their feasibility in real-world privacy-protection settings.

Chapter 3

Incrementality Representation Learning: Synergizing Past Experiments for Intervention Personalization

[1]

3.1 Introduction

Businesses are increasingly adopting personalization strategies to enhance their interactions with customers. *Targeting* and *tailoring* have emerged as critical pillars for crafting marketing interventions that enhance customer value. Targeting involves identifying specific customers most likely to respond favorably to interventions and exclusively targeting them. Tailoring further complements these efforts by modifying various design features of interventions to suit different customers, thereby maximizing their effectiveness.

At the core of effective targeting and tailoring is a firm's ability to accurately predict the *incremental impact* of various marketing interventions on individual customers. This precision is typically achieved through *conditional average treatment effect* (CATE) modeling, which estimates the incremental impact of specific interventions on customers based on their observed characteristics.

[1]Coauthored with Eva Ascarza and Ayelet Israeli.

This process traditionally consists of three stages. Initially, a randomized controlled experiment is conducted where a variety of predefined marketing interventions are assigned to a subgroup of customers. Subsequently, a CATE model is constructed for each intervention to predict its incremental impact using pre-treatment customer covariates. Finally, for a new set of customers, the intervention predicted to have the highest CATE is selected for each individual. This method has proven effective across various marketing domains, including free trial design [216], product experience [214], communication [64], and promotion [184, 223, 51]. Indeed, many companies have adopted this strategy, conducting experiments for every intervention of interest and developing CATE models to optimize targeting for each specific intervention.

However, this approach becomes impractical when interventions have many customizable design features. For example, optimizing ten features, each with three levels, would require testing $3^{10} = 59,049$ unique interventions in a full factorial design. In addition, each intervention would need a large customer base for accurate CATE estimation, making such experiments unfeasible or expensive for many organizations. This research introduces a novel approach that enables companies to target and tailor new interventions by *leveraging past experiments with readily available data*, instead of testing numerous interventions simultaneously [e.g., 65]. By pooling knowledge from past experiments, firms can reduce the need for extensive testing and cut costs while effectively targeting and tailoring their marketing interventions. Furthermore, synthesizing information across multiple experiments improves CATE prediction accuracy and targeting effectiveness, providing a significant advantage over analyzing each intervention in isolation, which often fails to detect treatment effect heterogeneity for personalization.

Leveraging past experiments to target and tailor new interventions presents several challenges. First, past experiments usually cover a limited range of interventions, complicating the generalization of model predictions to new, untested interventions. Besides, these experiments are typically tested on specific target populations, limiting the model's ability to predict impact across broader customer segments. Second, the complex interplay between high-dimensional design features and customer covariates makes it challenging to predict and interpret heterogeneous treatment effects across interventions, complicating decision-making when targeting and tailoring new interventions.

To address these challenges, we propose a novel causal machine learning framework called

Incrementality Representation Learning (IRL). By leveraging past experiments, this framework serves as a foundation model [29] for treatment effect prediction by integrating intervention design features and customer covariates from various experiments. This approach addresses the *cold-start problem* — evaluating new interventions without prior data. Additionally, our model condenses high-dimensional design features and customer covariates into *low-dimensional representations* that are highly predictive of treatment effects *across interventions*. This enhances the model's generalization to new interventions and broadens its applicability to previously unconsidered customer segments, even with small sample sizes and limited design space coverage.

The IRL methodology consists of two stages. First, we generate an unbiased proxy of the CATE for each individual in each experiment using the cross-fitted doubly robust score [126]. Next, we develop a specialized *deep multi-task representation learning* model. This model has two components: separate encoding networks for intervention design features and customer characteristics, transforming high-dimensional variables into low-dimensional latent representations that capture generalizable information on treatment effect heterogeneity; and a unified prediction network that uses these representations to predict the proxy CATE across all past experiments. This approach allows firms to predict treatment effects for interventions that differ from those previously tested but share similar characteristics, enabling the design and targeting of personalized interventions without extensive factorial experiments. Both the encoding and prediction networks are meticulously calibrated to enhance the accuracy of proxy treatment effects, ensuring their generalizability across various interventions and customer segments. We provide theoretical guarantees for the IRL framework and illustrate the accuracy benefits of the proposed deep learning architecture through theoretical analysis.

We assess the effectiveness of the proposed IRL model in the context of promotional campaigns for consumer packaged goods (CPG). By analyzing data from 274 experiments involving 6,884,833 customers on a leading customer engagement platform, we show that the IRL model significantly outperforms traditional approaches, which typically generate separate models for each intervention. Our model surpasses existing benchmarks in accurately predicting treatment effects and optimizing targeting for both tested promotional offers on *new customers* within the same target population and *untested offers* across different target segments. Notably, the IRL model even outperforms state-of-the-art CATE models trained on *actual* data collected from experiments

that tested these new interventions and customer segments (data purposefully excluded when training the IRL model). The results highlight the IRL model's capacity to effectively predict treatment effects by synergizing past experiments and generalizing successfully to new, untested settings.

In addition to enhancing targeting efficiency, we provide a practical framework for more effective intervention tailoring. We demonstrate how managers can identify critical design features and customer covariates through cluster analyses of learned incrementality representations, offering deeper insights into treatment effect heterogeneity compared to traditional analyses. This approach narrows the decision space by focusing on the most important aspects for tailoring interventions. We then introduce the *Segment Incrementality Plot* (SIP), which graphically illustrates how modifying key design feature affect treatment effects across customer segments. Combining the advantages of partial dependence [77] and individual conditional expectation plots [84], the SIP provides clear, segment-specific insights into promotion responsiveness and effectiveness. Our empirical analysis shows that optimal promotion design varies significantly across customer segments based on engagement levels, highlighting opportunities to boost profitability through tailored promotions

Our research makes three key contributions. Methodologically, we introduce a novel causal machine learning framework that combines deep multi-task representation learning with causal inference to predict treatment effects across different experiments, addressing cold-start and generalization issues for untested interventions or customer segments. We provide theoretical guarantees for the model and outline practical considerations for applying it to new interventions or customers. Substantively, we demonstrate that leveraging data from past experiments can significantly improve both targeting performance and tailoring capabilities, highlighting the potential to capitalize on underutilized past data. Managerially, we streamline the design of personalized marketing interventions with a decision-support tool that uses insights from past experiments, enhancing business outcomes by tailoring interventions to diverse customer segments.

The remainder of the paper is organized as follows: Section 3.2 provides a review of the literature. Section 3.3 introduces the IRL modeling framework and its theoretical foundations. Section 3.4 describes the empirical context and the data used in our study. Section 3.5 validates the proposed IRL method for treatment effect prediction and targeting for existing promotional offers,

and compares its performance against common modeling alternatives. Section 3.6 demonstrates the superior capability of the proposed IRL in targeting new interventions and customers, as well as tailoring interventions across different customer profiles. Section 3.7 concludes with a discussion of the findings and outlines directions for future research.

3.2 Related Literature

Our paper contributes to multiple streams of literature. First, it expands the field of marketing customization, which focuses on predicting purchase or click probabilities and identifying optimal offers based on behaviors without interventions [e.g., 7, 8, 216, 81, 140]. This research primarily aims to match relevant offers with customers rather than creating incremental impact. Recent studies have examined the causal effects of customizable components [201, 64, 51], but they often focus on simplified decision spaces and estimate the average treatment effect (ATE) across basic customer segments using linear models. Our research introduces a flexible machine learning framework to estimate heterogeneous causal effects across a wide range of interventions and customers, incorporating high-dimensional design features and customer covariates. In addition, unlike previous research that treats intervention design as a mathematical optimization problem [e.g., 7, 221, 222, 86], our model interpretation framework provides tools for managers to explore complex interactions between design features and customer covariates, enhancing their ability to understand the impact of specific design features on treatment effects.

Second, our research contributes to the evolving field of heterogeneous treatment effect estimation and targeting. While much of the methodological research [e.g., 199, 155, 180, 126] focuses on single experiments, our study expands these efforts to multiple interventions. Recent literature has explored treatment effect estimation for multiple treatments [45, 51, 64, 214] within relatively small intervention and customer spaces. [65] demonstrates the value of incorporating contextual embeddings from large language models to predict treatment effects of untested but similar email campaigns, using leave-one-email-out validation. Our paper extends this literature by (i) proposing a new model that handles complex, high-dimensional design features and customer covariates, improving generalization for untested interventions and customer segments, (ii) demonstrating the power of synergizing past experiments to predict treatment

effects for significantly different interventions (new product categories) and customer segments (new demographics for interventions of interest), and (iii) characterize the practical limitations in predicting treatment effects for novel treatments using past experiments.

Third, our work relates to the literature on the transportability of causal effects [184, 52]. This issue arises when the target population extends beyond the experimental population, making it challenging to identify causal effects due to the non-overlap of populations. Strategies to address this include (i) trimming the target population to match the experimental data coverage [e.g., 49, 161], (ii) integrating behaviors for non-treated segments from other studies with existing experimental data [e.g., 89, 228], and (iii) extrapolating using smoothness assumptions on CATE functions [153, 128, 166, 226, 65]. Our methodology combines the latter two strategies, exploiting the smoothness of the CATE function to predict treatment effects for previously untested interventions or customers by synergizing past experiments. We also provide a theoretical characterization of the transportability problem using domain adaptation theories [27].

Fourth, our research adds to the literature on multi-task and transfer learning [e.g., 227, 224], which concentrates on leveraging knowledge from diverse machine learning problems (source tasks) to enhance performance in specific focal tasks (target tasks). While there is growing interest in applying these methods to address business decision challenges, our contributions are distinct. Unlike prior studies that primarily focus on non-experimental settings to predict outcomes [132, 23, 209], our framework enables firms to synergize multiple randomized controlled experiments to predict causal effects. Furthermore, previous research constructs models individually for each of the past experiments and integrates these models with the actual data from the focal experiment [191]; our approach demonstrates the advantage of modeling all past experiments simultaneously, even *without conducting the focal experiment*. This benefit stems from our model's ability to directly share information across previous experiments, enabling companies to design new interventions without conducting an actual experiment.

Finally, this paper relates to the growing body of work on representation learning within the context of marketing. [55] introduce a multimodal representation learning framework for integrating diverse data types related to logos, while [80] and [35] propose extend the Word2Vec model [148] to generate latent product representations based on co-purchasing patterns. [81] and [35] extend these product representations into customer choice models for marketing mix

modeling. Unlike previous work, our framework prioritizes distilling critical information related to treatment effect heterogeneity instead of merely summarizing the original raw data (e.g., logo images or product co-purchase matrices) into dense vectors. Therefore, we utilize *supervised representation learning*, enabling our algorithm to extract the most generalizable information and exclude idiosyncratic details.

3.3 Model

In this section, we first specify the necessary conditions for utilizing past experiments to identify the treatment effect of a marketing intervention on a customer, taking into account both the design features of the intervention and the customer's covariates. We then discuss the potential challenges of generalizing past experiments to new settings and propose the IRL method as a solution. This approach enables companies to effectively tackle two crucial questions in designing personalized interventions: "whom to target" (the *targeting* problem) and "what to offer" (the *tailoring* problem) for new interventions

In line with our empirical application, we use "offers" to refer to interventions and "customers" as the unit of analysis. However, the applicability of our framework extends beyond marketing scenarios. It is relevant in any context where researchers or decision-makers have administered various treatments or stimuli to diverse units or individuals and aim to either enhance the efficacy of existing treatments or develop new ones.

3.3.1 Problem Setup

We consider a scenario where a company has already executed a series of experiments on a variety of offers. In each of these experiments, a distinct offer $o \in \{1, \cdots, O\}$, characterized by its design features $\mathbf{Z}_o \in \mathcal{Z} \subseteq \mathbb{R}^D$), was tested on a representative sample ($\mathcal{C}_o^E$) before being deployed to the target population ($\mathcal{C}_o$). In the experimental sample, customers were randomly assigned to either a treatment group that received the offer ($W_{o,i} = 1$) or a control group that did not ($W_{o,i} = 0$). For each participant i in $\mathcal{C}_o^E$, the company collects pre-treatment covariates $\mathbf{X}_{o,i} \in \mathcal{X} \subseteq \mathbb{R}^P$. These covariates include both static information, such as demographics, and offer-dependent variables, like previous purchase behaviors. We define $Y_{o,i}$ as the target outcome that the company aims to

influence through offer o, such as the total spending on items promoted by the offer.

Causal Estimand

When the objective is making targeting decisions for individual offers, the target causal estimand is typically the CATE for *each specific offer*. This is formally defined as follows:

$$\tau_o(\mathbf{X}_{o,i}) \equiv \mathbb{E}[Y_{o,i}(W_{o,i} = 1) - Y_{o,i}(W_{o,i} = 0)|\mathbf{X}_{o,i}],$$

where $Y_{o,i}(W_{o,i} = 1)$ and $Y_{o,i}(W_{o,i} = 0)$ denote the potential outcomes for customer i if they were to receive or not receive offer o, respectively. In this case, the company typically estimates O distinct CATE models based on experimental data and formulates separate targeting strategies for each offer.

While this estimand may be effective for targeting decisions on previously tested offers, estimating CATEs individually does not allow companies to effectively target new offers due to the absence of experimental data for these offers. A naive approach to address this issue might assume a uniform CATE function across all offers and construct a CATE model using data from all experiments. However, this approach does not account for how different design features impact incremental effects across various customer segments. Consequently, this model falls short in its ability to tailor offers to specific customer groups.

To effectively target and tailor new interventions, we propose using the CATE conditioned on both the design features and customer covariates as the target estimand for personalization. This estimand is defined as follows:

$$\tau(\mathbf{Z}_o, \mathbf{X}_{o,i}) \equiv \mathbb{E}_{\mathcal{C}_o}[Y_{o,i}(\mathbf{Z}_o) - Y_{o,i}(\mathbf{0})|\mathbf{Z}_o, \mathbf{X}_{o,i}], \tag{3.1}$$

where $Y_{o,i}(\mathbf{Z}_o)$ is the potential outcome for observation i should they receive an offer with design features $\mathbf{Z}_o$, and $Y_{o,i}(\mathbf{0})$ is the potential outcome in the absence of offer o.

By integrating data from multiple experiments, we can estimate (3.1) by leveraging variations in design features and customer covariates. This approach offers several advantages over estimating CATE models individually for each offer. First, analyzing design features across various offers helps develop new marketing strategies tailored to maximize incremental gains. Second, using

data from diverse offers and customer pools reveals treatment effects on previously ineligible individuals, enhancing the generalizability of predictions to new customer segments. Third, it improves targeting accuracy by considering interactions between offer designs and customer demographics. Finally, pooling data across experiments increases the sample size, improving the precision of treatment effect estimates.

Identification Assumptions

The ideal approach for estimating the quantity described in (3.1) would involve conducting an experiment with a full factorial design, which randomly assigns all possible combinations of design features across a large customer base. However, this approach poses significant operational challenges, particularly when it requires the random allocation of many potential treatments simultaneously — approximately one billion possible offers in our empirical application — making it impractical for most scenarios. Therefore, our goal is to estimate (3.1) using readily available data, specifically by leveraging experimental data from offers that have already been tested.

Since this approach does not involve direct manipulation of each offer's design features, it is essential to ensure that past experiments meet specific conditions to identify the target estimand in (3.1) using data from multiple single-offer experiments. Below, we outline the assumptions required to ensure identification based on past experiments.

Assumption 3.1 *The following identification assumptions are made:*

1. (Overlap) *In each experiment* $o \in 1, \cdots, O$*, the assignment of treatment is subject to random variation, that is,* $0 < \mathbb{P}(W_{o,i} = 1 | \mathbf{X}_{o,i}) < 1, \quad \forall\, \mathbf{X}_{o,i} \in \mathcal{X}_o, \;\; o = 1, \cdots, O.$

2. (Unconfoundedness) *In each experiment* $o \in 1, \cdots, O$*, the treatment assignment is free from unobserved confounders, that is,* $\{Y_{o,i}(\mathbf{Z}_o), Y_{o,i}(\mathbf{0})\} \perp W_{o,i} \mid \mathbf{X}_i, \mathbf{Z}_o, \quad \forall i \in \mathcal{C}_o^E, \; \forall o = 1, \cdots, O.$

3. (No Interference) *The potential outcomes for customer i in response to offer o are independent of the presence of other offers received by customer i and offers received by other customers i′, meaning that* $\{Y_{o,i}(\mathbf{Z}_o), Y_{o,i}(\mathbf{0})\} \perp \{W_{o',i}\}_{o' \neq o}$ *and* $\{Y_{o,i}(\mathbf{Z}_o), Y_{o,i}(\mathbf{0})\} \perp \{W_{o',i'}\}_{i' \neq i,\; o' = 1, \cdots, O}$.

4. (Stability) *The experimental sample (including both the treatment and control groups) and the future target population are comparable, that is,* $\mathbb{E}_{\mathcal{C}_o^E}[Y_{o,i}(\mathbf{Z}_o) | \mathbf{X}_i] = \mathbb{E}_{\mathcal{C}_o}[Y_{o,i}(\mathbf{Z}_o) | \mathbf{X}_i]$ *and* $\mathbb{E}_{\mathcal{C}_o^E}[Y_{o,i}(\mathbf{0}) | \mathbf{X}_i] =$

$\mathbb{E}_{\mathcal{C}_o}[Y_{o,i}(\mathbf{0})|\mathbf{X}_i]$ *for any* $o \in \{1, \cdots, O\}$,

The first and second assumptions are the standard overlap and unconfoundedness assumptions in the literature [117]. The third assumption ensures that the behavior of one customer does not influence another, and that individual offers do not interfere with each other. In our empirical setting, the likelihood of offer interference is minimal because: (i) customers did not receive offers for identical items simultaneously; (ii) the outcome variable solely captures purchases of items promoted by the focal offer; and (iii) both treatment and control groups had an equal likelihood of receiving other offers, due to the company's implementation of complete randomization. The fourth assumption ensures that the future targeting pool for an offer will behave in the same manner as the customers included in the experiment.

Under these assumptions, the customer response function $\mathbb{E}[Y_{o,i}(\mathbf{Z})|\mathbf{X}_{o,i}]$ is identified at $\mathbf{Z} = \mathbf{Z}_o$ and at $\mathbf{Z} = \mathbf{0}$ for customers within $\mathcal{C}_o$. We can express this result as:

$$\begin{aligned} \mathbb{E}_{i\in\mathcal{C}_o}[Y_{o,i}(\mathbf{Z}_o)|\mathbf{X}_{o,i}] &= \mathbb{E}_{i\in\mathcal{C}_o^E}[Y_{o,i}|W_{o,i}=1,\mathbf{X}_{o,i}], \quad \forall\ \mathbf{X}_{o,i}\in\mathcal{X}_o,\ \ o=1,\cdots,O, \\ \mathbb{E}_{i\in\mathcal{C}_o}[Y_{o,i}(\mathbf{0})|\mathbf{X}_{o,i}] &= \mathbb{E}_{i\in\mathcal{C}_o^E}[Y_{o,i}|W_{o,i}=0,\mathbf{X}_{o,i}], \quad \forall\ \mathbf{X}_{o,i}\in\mathcal{X}_o,\ \ o=1,\cdots,O. \end{aligned} \tag{3.2}$$

This formulation implies that one can aggregate data from prior experiments and employ well-established CATE models to estimate the treatment effect given specific design features and customer covariates.

Interpolation and Extrapolation Challenge

It is crucial to recognize that Equation (3.2) provides a point-wise identification result, which only guarantees the identification of the CATE for those combinations of design features and customer covariates that have been previously observed. As such, predicting (3.1) for *untested* offers and customer segments — key to our research objectives — requires some degree of interpolation or extrapolation from the CATEs of previously tested offers.[2] Given the high-dimensional nature of design features and customer covariates, it is inevitable that some combinations of offer design and customer segments will remain unobserved, despite numerous past experiments [20].

Since the most widely-used CATE models, such as the R-learner [155] and Causal Forest

[2]This challenge is common in many causal inference settings, particularly when dealing with a finite experimental sample and high-dimensional customer covariates that influence treatment effect heterogeneity [161, 101].

[199], are primarily designed for scenarios within a single intervention and do not support the interpolation and extrapolation of CATEs from past experiments to untested interventions, there is a pressing need to develop a model that can effectively generalize CATE predictions to new combinations of designs and customers. Such a model should not only predict the CATE function for untested offers and customers using design features and covariates but also possess robust generalization ability by learning the common structure across different offers while discarding idiosyncratic, offer-specific patterns.

Solution Concept

To tackle this challenge, we introduce the IRL framework, which merges causal machine learning [180, 155, 126] with multi-task representation learning [26, 28, 145, 205]. This approach aims to accomplish two primary objectives: firstly, to derive low-dimensional representations from design features and customer covariates that summarize generalizable information regarding treatment effect heterogeneity across experiments; and secondly, to learn the common relationship between these representations and the actual treatment effects across different offers using a unified prediction model.

The proposed model architecture provides an accurate foundation model for treatment effect predictions for three reasons. First, extracting low-dimensional common representations across all promotional offers can often mitigate the risks associated with extrapolation in the high-dimensional spaces of design features and customer covariates [20, 52]. Typically, modeling in such high-dimensional decision spaces demands a significantly larger sample size to achieve adequate coverage. However, if there is an underlying low-dimensional latent space of input variables that effectively captures the essential information of CATE, achieving sufficient coverage in this reduced space becomes much more feasible [21, 30].

Second, by training a single prediction model to minimize prediction errors across all experiments, we ensure that IRL discards offer-specific idiosyncrasies in the relationships between the learned representations and the true CATE. This approach enhances generalizability to untested offers and customer segments by identifying and leveraging overarching patterns of treatment effect heterogeneity across experiments.

Third, the proposed IRL model enhances the accuracy of CATE predictions by increasing statistical power through the pooling of information across experiments. Leveraging knowledge from multiple prediction tasks consistently outperforms models trained on individual tasks alone, as supported by both theoretical [26, 28, 145] and empirical research [191, 196, 224]. In real-world marketing applications, estimating heterogeneous treatment effects for a single experiment can be challenging due to small treatment effects being overshadowed by significant noise in the outcome variable [112]. By pooling information from multiple experiments, IRL effectively increases the effective sample size and, thus, the statistical power, enhancing prediction accuracy compared to models that treat each experiment independently.

Next, we detail the proposed IRL model, present its theoretical foundations, and discuss practical considerations for its application to real-world data.

3.3.2 Incrementality Representation Learning

Our proposed IRL framework consists of two stages. First, we use state-of-the-art techniques to construct an unbiased proxy for the true CATE. Following this, we develop a deep supervised representation learning model tasked with predicting this proxy CATE. The model employs design features and customer covariates to generate low-dimensional representations that capture generalizable insights about treatment effect heterogeneity. These representations are then used to predict the proxy CATE.

Derive Unbiased Proxy for CATEs

The first step creates an unbiased proxy for the CATE corresponding to each customer and offer. This proxy will then act as the prediction target for our deep multi-task representation learning algorithm. The following theorem demonstrates how to derive an unbiased proxy for the CATE defined in (3.2) using the doubly robust score [126].

Theorem 3.1 (Unbiased Score for CATE) *Let* $\mu_{o,w}(\mathbf{X}_{o,i}) = \mathbb{E}_{\mathcal{C}_o^E}[Y_{o,i}|W_{o,i} = w, \mathbf{X}_{o,i}]$ *represent the conditional expected outcome for offer* o*, and let* $\pi_o(\mathbf{X}_{o,i}) = \mathbb{P}_{\mathcal{C}_o^E}[W_{o,i} = 1|\mathbf{X}_{o,i}]$ *denote the propensity score of being treated in experiment* o*. Then, under Assumption 3.1, we have the following identification result for the doubly robust score:*

$$\tau(\mathbf{Z}_o, \mathbf{X}_{o,i}) = \mathbb{E}_{\mathcal{C}^E}\left[\mu_{o,1}(\mathbf{X}_{o,i}) - \mu_{o,0}(\mathbf{X}_{o,i}) + \frac{W_{o,i} - \pi_o(\mathbf{X}_{o,i})}{\pi_o(\mathbf{X}_{o,i})[1-\pi_o(\mathbf{X}_{o,i})]}\left[Y_{o,i} - \mu_{o,W_{o,i}}(\mathbf{X}_{o,i})\right] \middle| \mathbf{X}_{o,i}, \mathbf{Z}_o\right]$$

for all $\mathbf{X}_{o,i} \in \mathcal{X}_o,\ o = 1, \cdots, O$. *Here,* $\mathcal{C}^E = \bigcup_{o=1}^{O} \mathcal{C}_o^E$ *denotes the collection of all observations from past experiments.*

The proof of Theorem 3.1 is provided in C.2.2. Theorem 3.1 demonstrates that to predict the CATE defined in (3.1), one can first calculate the doubly robust score for each offer, then combine these scores across offers, and finally build a machine learning model using the design features and covariates across offers to predict the score.

In our empirical application, we utilize cross-fitting [155] to construct the doubly robust score:

$$\widetilde{\tau}_{o,i} \equiv \left[\widehat{\mu}_{o,1}^{[-i]}(\mathbf{X}_{o,i}) - \widehat{\mu}_{o,0}^{[-i]}(\mathbf{X}_{o,i})\right] + \frac{W_{o,i} - \widehat{\pi}_o^{[-i]}(\mathbf{X}_{o,i})}{\widehat{\pi}_o^{[-i]}(\mathbf{X}_{o,i})[1-\widehat{\pi}_o^{[-i]}(\mathbf{X}_{o,i})]}\left[Y_{o,i} - \widehat{\mu}_{i,W_{o,i}}^{[-i]}(\mathbf{X}_{o,i})\right], \tag{3.3}$$

where $\widehat{\mu}_{o,w}^{[-i]}(\mathbf{X}_{o,i})$ represents a (machine learning) model that predicts the expected outcome $Y_{o,i}$ for experiment o given the treatment assignment $W_{o,i} = w$ and the covariates $\mathbf{X}_{o,i}$, and $\widehat{\pi}_o^{[-i]}(\mathbf{X}_{o,i})$ denotes the propensity score model for experiment o, estimating the probability of receiving the treatment based on the covariates. The superscript $[-i]$ indicates that both the expected outcome and the propensity score predictions for observation i are based on models trained on data excluding that of observation i.

Deep Learning Framework for IRL

Next, we develop a deep multi-task representation learning model for CATE prediction. This model utilizes two distinct *encoding networks*: $\mathbf{\Phi} : \mathbb{R}^D \to \mathbb{R}^{K_1}$ for design features, where $K_1 < D$, and $\mathbf{\Psi} : \mathbb{R}^P \to \mathbb{R}^{K_2}$ for customer covariates, where $K_2 < P$. These networks convert the high-dimensional design features and customer covariates into low-dimensional representations: $\boldsymbol{\zeta}_o = \mathbf{\Phi}(\mathbf{Z}_o)$ for design features and $\boldsymbol{\chi}_{o,i} = \mathbf{\Psi}(\mathbf{X}_{o,i})$ for customer covariates. The primary function of these encoding networks is to extract the most relevant and generalizable information for predicting treatment effects across different offers. Then, the model implements a prediction network, $\tau : \mathbb{R}^{K_1+K_2} \to \mathbb{R}$, which utilizes these condensed representations to predict the doubly robust score (as detailed in Section 3.3.2) for each observation within an experiment.

Model Architecture. Figure 3.1 illustrates the proposed model architecture for IRL. Here, the model first transforms the high-dimensional design features $\{\mathbf{Z}_o : o = 1, \cdots, O\}$ and customer covariates $\{\mathbf{X}_{o,i} : \ i \in \mathcal{C}_o^E, \ o = 1, \cdots, O\}$ into lower-dimensional representations ($\boldsymbol{\zeta}_o$ for design features and $\boldsymbol{\chi}_{o,i}$ for customer covariates) through two separate deep neural networks ($\boldsymbol{\Phi}$ for design features and $\boldsymbol{\Psi}$ for customer covariates). Next, these representations are concatenated into extended vectors and inputted into another deep neural network (τ) to predict the proxy CATE ($\widetilde{\tau}_{o,i}$).

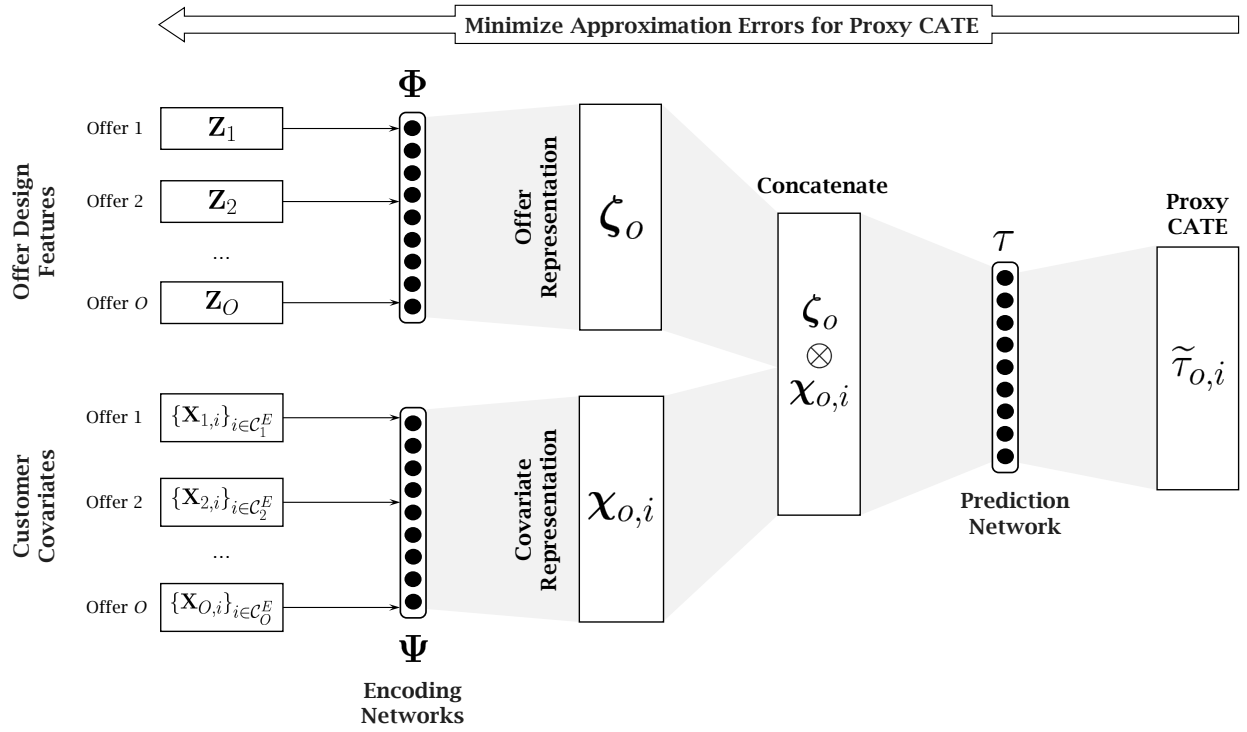


Figure 3.1: *Proposed Model Architecture for Incrementality Representation Learning*

One key design choice in our proposed architecture is the *separate construction* of representations for design features and customer covariates, instead of combining them into a single shared representation. This approach provides two critical benefits. First, it reduces the risk of overfitting and enhances generalization. By isolating the representations, this method prevents idiosyncratic interactions between design features and customer covariates that are not generalizable across offers during the representation learning process. The theoretical benefits of this approach are detailed in C.1.3.

Second, separating design features and customer covariates into distinct representations

enhances the model's interpretability. As detailed in Section 3.6.2, clustering design feature representations identifies distinct offer groups with similar treatment effect patterns and highlights influential design features. Similarly, clustering customer covariate representations detects customer segments with similar responses, improving targeting strategies. In contrast, a combined representation might obscure critical insights into how specific aspects of the offer or customer characteristics influence treatment effects, thus complicating the analysis for managers.

Another key design aspect of the proposed model architecture is the decision to *share all parameters* across the prediction network for different experiments, rather than creating offer-specific prediction networks for each individual offer — a common approach in multi-task learning. This decision is driven by two considerations. First, the model is designed to predict treatment effects for *untested* offers, for which no experimental data exists. Offer-specific networks would not be feasible in this scenario, as they would be unable to predict the impact of previously untested offers. Second, from an accuracy perspective, the CATE signal $\widetilde{\tau}_{o,i}$ often exhibits significant variance, primarily due to inverse propensity score weighting and unobserved customer heterogeneity [112]. Constructing an accurate prediction network for each offer is challenging due to the high level of noise and the relatively small sample sizes typical of individual experiments. Conversely, a unified prediction network that consolidates all experiments can enhance prediction stability by leveraging a larger training dataset.

Model Calibration. We calibrate the three neural networks — $\mathbf{\Phi}$, $\mathbf{\Psi}$, and τ — simultaneously, ensuring that $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are optimized to capture the most salient representations for treatment effect prediction. Specifically, we first define the architecture for each network. Following this, we compile data from all experiments and calibrate $\mathbf{\Phi}, \mathbf{\Psi}, \tau$ by minimizing the squared error between the model's predictions and the doubly robust scores:

$$\underset{\tau, \mathbf{\Phi}, \mathbf{\Psi}}{\arg\min} \sum_{o=1,\cdots,O} \sum_{i \in C_o^E} w_o \left[\tau(\mathbf{\Phi}(\mathbf{Z}_o), \mathbf{\Psi}(\mathbf{X}_{o,i})) - \widetilde{\tau}_{o,i}\right]^2, \tag{3.4}$$

where w_o denotes the relative importance weight assigned to each offer, which can be manually specified or treated as tuning parameters. For simplicity, we pick $w_o = 1$ for all offers in our empirical application.

Theoretical Guarantees

In C.1, we provide theoretical guarantees for the proposed IRL framework and model architecture. First, we establish upper bounds on the prediction error for the true CATE for any IRL model, drawing on the theoretical results from [193]. We detail how the error bound in our context is determined by three key factors: (i) the number of offers previously tested, (ii) the sample size of each past experiment, and (iii) the complexity of the representation and prediction models. Next, we illustrate how our specialized deep learning architecture can improve these theoretical upper bounds compared to traditional deep learning models. Specifically, we demonstrate the benefits of dimension reduction in deep representation learning, the advantages of employing distinct network structures for modeling offer and customer representations, and the benefit of learning shared parameters in the prediction network. This theoretical foundation supports the enhanced accuracy of our approach over conventional deep learning architectures in our empirical applications.

3.3.3 Practical Considerations

The IRL framework is a powerful tool for leveraging insights from past experiments to improve targeting and personalization for both existing and new offers. However, its effectiveness for untested offers and customers hinges on the similarity between the data-generating processes of the past experimental data and the new dataset. When managers apply models trained on past experimental data, there are two considerations that require particular attention. First, if past experiments lack diversity in offer design or customer covariates, a CATE model may not generalize well to new offers or customer groups significantly different from those in previous scenarios. Second, if customer responses to offers evolve over time or untested customers have different sensitivities (i.e., the stability assumption in Assumption 3.1 is violated), reliance solely on historical data may cause the CATE model to generate inaccurate predictions for new data.

To formalize this, we denote the data-generating process for $\mathbf{Z}_o, \mathbf{X}_{o,i}$ in past experiments as $\mathcal{D}_{\text{past}}$, with the true CATE for these offers and customers represented by τ_{past}. The managers aim to predict the treatment effects for a new set of offers and customers, characterized by a different data-generating process $\mathcal{D}_{\text{new}}$ for their design features and customer covariates, with their true

CATE function expressed as τ_{new}. Building on classical domain adaptation theory [27], we identify three critical factors that influence the generalization performance of any CATE model trained on past experiment data, as outlined in the following theorem.

Theorem 3.2 (Generalization Error) *Let $\widehat{\tau}$ be a CATE model trained on past experiments by the empirical risk minimization problem:*

$$\min_{\widehat{\tau}} \quad \frac{1}{\sum_{o=1}^{O} |\#\mathcal{C}_o^E|} \sum_{o=1}^{O} \sum_{i \in \mathcal{C}_o^E} \ell\left(\widehat{\tau}(\mathbf{Z}_o, \mathbf{X}_{o,i}), \widetilde{\tau}_{o,i}\right),$$

where $\widetilde{\tau}_{o,i}$ is an unbiased estimate of $\tau_{\text{past}}(\mathbf{Z}_o, \mathbf{X}_{o,i})$, ℓ is the squared loss function, and $|\#\mathcal{C}_o^E|$ denotes the number of observations in $\mathcal{C}_o^E$. Also, assume that $\widehat{\tau}$, τ_{past}, and τ_{new} are bounded. Then, the expected loss for the new offers can be bounded as follows:

$$\mathbb{E}_{\mathcal{D}_{\text{new}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{new}}\right)\right] \leqslant \underbrace{\mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{past}}\right)\right]}_{\textit{Prediction Error for Past Experiments}} + C \underbrace{\delta(\mathcal{D}_{\text{past}}, \mathcal{D}_{\text{new}})}_{\textit{Attribute Shift}} + \underbrace{\min\left\{\mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\tau_{\text{past}}, \tau_{\text{new}}\right)\right], \mathbb{E}_{\mathcal{D}_{\text{new}}}\left[\ell\left(\tau_{\text{past}}, \tau_{\text{new}}\right)\right]\right\}}_{\textit{Concept Shift}},$$

where C is a constant, $\delta(\mathcal{D}_{\text{past}}, \mathcal{D}_{\text{new}})$ is the total variation distance between the two data generating distributions, i.e., $\delta(\mathcal{D}_{\text{past}}, \mathcal{D}_{\text{new}}) = \sup_{B \in \mathcal{B}} \left|\mathbb{P}_{\mathcal{D}_{\text{past}}}(B) - \mathbb{P}_{\mathcal{D}_{\text{new}}}(B)\right|$, and $\mathcal{B}$ denotes the collection of all Borel-measurable sets applicable to $\mathcal{D}_{\text{past}}$ and $\mathcal{D}_{\text{new}}$. For simplicity, $(\mathbf{Z}_o, \mathbf{X}_{o,i})$ are omitted from each function's notation.

The proof of Theorem 3.2 is provided in C.2.5. Theorem 3.2 identifies three critical factors influencing the upper bound of generalization error: (i) the accuracy of the model's predictions for existing offers, (ii) the differences in the distribution of design features and customer covariates between past and new offers (i.e., *attribute shifts*), and (iii) the disparity in the true CATE functions between past and new offers (i.e., *concept shifts*). The first factor represents prediction error when new data's distribution and true CATE function match those of past data. The second factor addresses challenges when new offers have design features and customer covariates that differ significantly from past experiments. For instance, if past experiments lacked offers for baby products but the new set includes them, this can hinder generalization. The third factor highlights potential errors when the responses of untested customers differ significantly from those

previously tested, such as household versus non-household customer sensitivities to promotions.

In our empirical analysis, we demonstrate that the proposed IRL model outperforms existing methods in generalizing to previously untested offers, specifically in product categories that had never been tested before, and to customer segments that were previously ineligible for certain offers (see Section 3.6.1). However, the model's ability to generalize to untested offers or customer segments that are substantially different from those previously tested remains an empirical question. If companies are seriously concerned about these issues, they could conduct small-scale experiments with the new offers and customers, then apply the proposed IRL model to this new experimental data. Alternatively, they could perform model fine-tuning to enhance the performance of the IRL model trained on past experimental data.

3.4 Empirical Context and Data

3.4.1 Empirical Context

The dataset is provided by a leading consumer-engagement platform in North America. Customers on this platform earn reward points by scanning their shopping receipts, which they can then redeem for gift cards or products. The platform generates revenue by providing customer relationship management services to CPG companies and retailers, who pay commissions for these services. A key feature of the platform is its assistance in creating *special promotional offers* for these partners, such as "buy two bottles of wine from Brand W and get 2,000 points." These offers are prominently displayed on the platform's discovery page, allowing users to earn additional points by submitting receipts that meet each offer's requirements.

The primary goal of the platform is to boost the incremental sales of promoted products. With numerous partners aiming to expand their customer base and increase sales through branded promotional campaigns, the company seeks to deliver more targeted offers on the discovery page to reduce the volume of irrelevant offers. Furthermore, since the platform's value proposition for partners centers on demonstrating a sales lift, it is crucial to strategically allocate limited promotional budgets to customers most likely to generate the highest incremental sales. Consequently, the success of the platform depends on precisely tailoring and targeting optimal promotional offers to the audiences most likely to make additional purchases in response to these

promotions.

3.4.2 Data Description

To demonstrate sales lift, the platform routinely conducts randomized controlled experiments for its special promotional offers. For each offer, about 10% of eligible customers are randomly chosen to form a control group that does not receive the offer. We analyzed 274 such experiments, each involving different promotions conducted between October 2022 and April 2023. Each of these promotions was tested on 10,000 to 200,000 customers.[3]

Offers

Each offer is characterized by several key design features: (i) the promoted product and its category, (ii) point-earning criteria, like minimum purchases or quantities, (iii) points awarded for fulfilling these criteria, and (iv) extra details like redemption limits or promotional tactics. Overall, our dataset comprises fourteen design features, including six numerical and eight discrete variables, expanding to 46 variables following dummy encoding. A summary of these features is detailed in Table 3.1.

Customers

Our sample contains information from 6,884,833 distinct customers, each of whom was included in at least one experiment. In total, we collect 40 pre-treatment covariates for capturing heterogeneous customer reactions to a promotional offer. Below is an overview of each covariate group:

1. **User attributes:** Customer characteristics such as gender, age, duration of membership on the platform (measured at the time they received the offer), and whether they joined through a referral.

[3]We focus exclusively on the 76% of offers that exhibit non-negative ATEs. Offers with negative effects are excluded as they may be influenced by substantial noise or issues with execution. However, our findings remain consistent even if we use all the offers, albeit with higher standard error in targeting performance. See C.6.2 for more details.

Table 3.1: *Summary of Offer Design Features.*

Discrete Design Features

Variable	Unique Value	Top Counts
Is the offer a multi-transaction offer?	2	False: 162, True: 112
Is the offer stackable?	2	False: 33, True: 241
Criteria: Purchase from selected items	2	False: 242, True: 32
Criteria: Meet minimum quantity	2	False: 34, True: 240
Criteria: Available for club members only	2	False: 260, True: 14
Criteria: Actions for reward claim	2	Quantity: 240, Spending: 34
Promoted product category	19	Beer: 57, Personal Care: 55, Baby: 52, Wine: 31
Promotion tactic	14	Dollar_or_unit: 125, Frequency: 57, Competitive_targeting: 47

Numerical Design Features

Variable	# Missing	Mean	S.D.	Min	P25	Median	P75	Max
Points awarded	0	2,587	2,305	350	1000	2,000	3,500	20,000
Minimum required spending ($) to redeem the reward points	240	18.9	11.4	5.0	10.0	15.0	30.0	50.0
Minimum required quantity to redeem for the reward	34	1.37	0.614	1	1	1	2	4
Total number of times the offer can be redeemed	0	1.84	1.5	1	1	1	2	10
Maximum redemptions allowed per transaction	0	1.41	1.11	1	1	1	1	5
Duration of the promotion (in days)	0	43.8	30.6	5	28	31	60	176

Note: For instance, the offer "Spend $30 on selected sizes of diapers and earn 3,000 points" is characterized as a single-transaction, non-stackable offer. It specifies the criteria as *purchase from selected items* along with a *spending* threshold for claiming the reward. The variable "Is the offer stackable?" indicates whether a purchase related to this offer can also count towards another offer. The variable "promotion tactics" describes the strategy behind the offer. For instance, "Dollar_or_unit" is designed to increase the dollars spent or units purchased of specified items, while "Frequency" aims to enhance the purchase frequency of the items.

2. **Non-contextual behaviors:** Summaries of purchase behaviors 90 days before an experiment, including recency (whether any receipt was uploaded 7, 30, or 60 days prior to the promotion), frequency (total number of uploaded receipts), and monetary metrics (average order value per receipt and average item count per receipt), as well as preferences for the product categories or merchants, defined by the proportion of receipts including a specific product category or those associated with a particular merchant.

3. **Contextual behaviors:** Purchase behaviors related to the promoted items 90 days prior to the experiment, including recency, frequency, and monetary metrics.

Experiments

In each of the 274 experiments, a specific promotional offer was tested on a target population (i.e., the eligible customer pool).[4] Eligible customers of each offer were randomly divided into

[4]At any point in time, customers are also exposed to a handful of other promotional offers not subject to any experimentation. The randomization process ensured that both the treated and control groups within an experiment were exposed to comparable non-tested offers, facilitating the identification of the focal offer's causal effect.

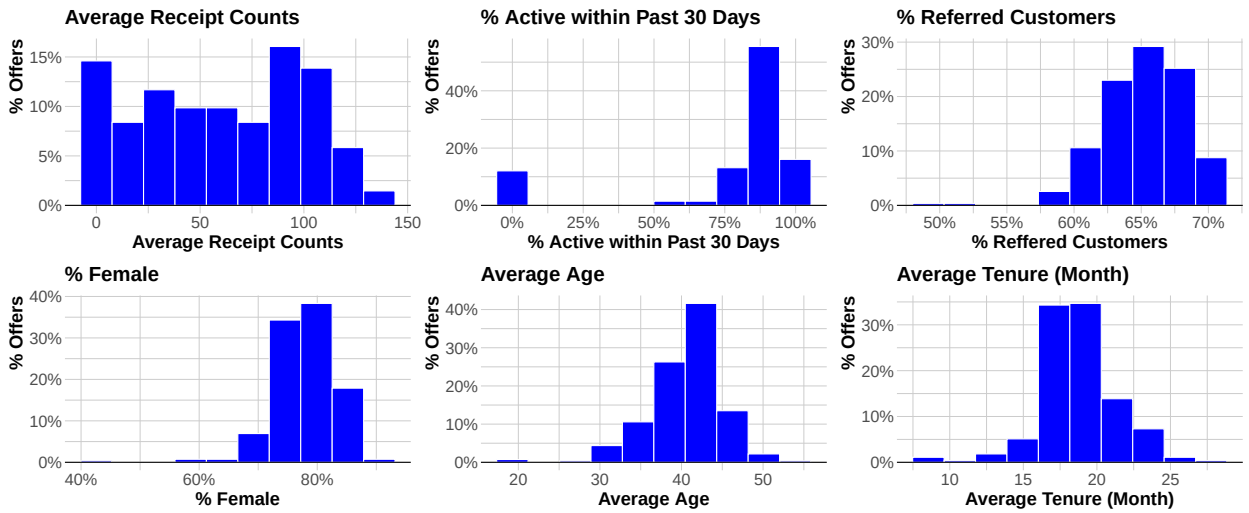


Figure 3.2: *Distribution of Average Covariate Value across the Eligible Customer Base*

treatment and control groups, with 10% of the participants assigned as control customers. These control customers did not receive the offer, allowing for an assessment of the promotion's causal impact. Detailed randomization checks are provided in C.3.1.[5]

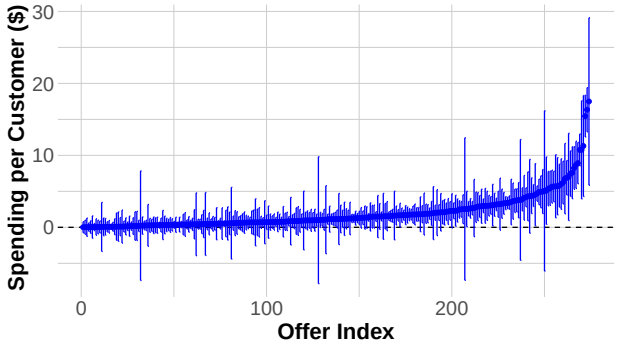
The focal company tests its offers on different target populations, resulting in diverse distributions of customer covariates across offers. Figure 3.2 displays the distribution of average values for selected covariates among the eligible customer base for each offer. For example, the top middle figure indicates that approximately 12% of the offers target dormant customers, defined as those who have not uploaded any receipts in the past 30 days. Meanwhile, the bottom middle figure reveals that around 10% of the offers specifically aim at younger customers, characterized by an average age of under 35. These patterns highlight the varying target populations for different offers.

3.4.3 Variations in Average Treatment Effects

The primary goal of the focal company is to increase total dollar spending on promoted items during the effective period of the offers. Therefore, this metric has been used as the outcome

[5]Although a small portion of customers received more than one tested offer during our observation window, we believe that the potential interference is minimal. First, on average, only 7% of customers received multiple tested offers simultaneously, and none of these offers promoted identical items. Second, both the treatment and control groups for the focal offer were equally likely to receive offers other than the focal one, reducing the likelihood of offer interference. In C.3.2, we provide empirical evidence demonstrating that the interference from multiple treatments is minimal.

variable for estimating treatment effects throughout this research. Figure 3.3 illustrates the ATE of each offer along with its 95% confidence interval. Using a two-tailed t-test at a 5% significance level, we find that 43% of the offers have a statistically significant positive impact on dollar spending on promoted products.



Note. Each point reports the average treatment effect of a promotional offer together with its 95% confidence interval.

Figure 3.3: *Average Treatment Effects of 274 Promotional Offers*

In C.7.1, we further explore how variations in ATEs correlate with observable characteristics. Overall, we find a significant positive correlation between some design features (such as reward points) and average values of customer covariates for each experiment (such as offer-related average order values). These results suggest the potential for the platform to utilize both design features and customer covariates to predict the treatment effects of promotional offers across customers.

3.5 Model Implementation and Predictive Validity

In this section, we detail the implementation of the proposed IRL model within this context and assess its performance relative to existing benchmarks in the literature.

3.5.1 Implementation of IRL

We implement the IRL model by calculating the doubly robust score and then developing a deep multi-task representation learning model to predict the score. (Please refer to C.4 for detailed model specifications described in this section.)

Doubly Robust Score Estimation

We calculate the doubly robust score as follows: For each promotional offer, we construct separate conditional outcome models for the treatment and control groups using the 40 pre-treatment covariates detailed in Section 3.4. We use XGBoost [37] with 10-fold cross-fitting to generate $\widehat{\mu}_{o,1}^{[-i]}$ and $\widehat{\mu}_{o,0}^{[-i]}$ in Equation (3.3), ensuring that the prediction error from these models does not introduce bias into the CATE estimation [40]. For the propensity score calculation, we use the proportion of customers treated in each experiment, leveraging the random treatment assignment. These predicted outcomes and propensity scores are then incorporated into Equation (3.3) to calculate the doubly robust score for each offer.

Network Architectures and Parameter Calibration

For the encoding networks, we utilize a five-layer fully-connected neural network architecture.[6] The input layers accommodate $D = 46$ dimensions for design features and $P = 40$ dimensions for customer covariates, respectively. Each network comprises three hidden layers, each containing 10 hidden units. The output layers of the networks produce K_1-dimensional and K_2-dimensional representations for design features and customer covariates, respectively. We set $K_1 = K_2 = 10$ because the dimensions of design features and customer covariates are similar. The Rectified Linear Unit (ReLU) activation function ($\sigma(x) = \max(0, x)$) is employed across all units. [7]

For the prediction network, we also employ a five-layer fully-connected neural network architecture. The input to this network is the concatenated representations from the design features and customer covariates, resulting in an input layer of 20 dimensions. Each of the hidden layers contains 10 hidden units. The network's output layer is tasked with generating the CATE prediction. We use the ReLU activation function for all hidden units and a linear activation function for the final prediction.

[6]We also tested shallower and deeper network architectures in C.6 and found that while the predictive accuracy is consistent across various network depths, targeting effectiveness is poorer with shallower architectures, although they still significantly outperform individual modeling.

[7]We find a similar targeting performance when choosing different values of K_1 and K_2 (see C.6 for results when choosing $\{10, 20, 30, 40, 50\}$ dimensions). The results indicate no significant differences in performance with varying representation dimensions up to 40, while the performance typically drops when $K_1 = K_2 = 50$. We also find that deeper neural networks tend to perform slightly better with fewer representation dimensions compared to higher dimensions.

We calibrate the incrementality representation network by optimizing the loss function detailed in Equation (3.4) through the adaptive moment estimation (Adam) algorithm [131]. This method is widely applied in deep learning for its ability in handling sparse gradients and adapting the learning rate for each parameter, facilitating efficient and effective model training. We also utilize mini-batch training to ensure computational efficiency and stabilize the training process.

3.5.2 Predictive Validity

To evaluate our model's performance relative to existing methods, we first assess its ability to accurately predict and correctly rank the treatment effects for each of the 274 offers. While this validation step is secondary to our primary objective of informing the targeting and tailoring of new offers, it aligns with established benchmarks in the field for evaluating different targeting models [9, 184, 218, 104, 112]. This assessment also shows that IRL can enhance targeting policies for existing offers on customers from the same populations.

Methods for Comparison

We evaluate the IRL model's performance by comparing it with two types of benchmarks.

Individual CATE Models. Our first benchmark involves developing *a distinct CATE model for each promotional offer*, using the doubly robust score computation method outlined in Section 3.5.1. We then utilize a deep neural network designed to predict these scores, excluding design features in the model since they are constant for each specific offer. This network consists of eight fully connected layers, with a total parameter count comparable to that of the network architecture described in Section 3.5.1.[8]

Joint CATE Models. Our second benchmark compares the proposed incrementality representation learning architecture to other joint CATE models trained using data from all 274 experiments. Specifically, we implement DR-learners using data pooled from these experiments [65]. All deep

[8]We also evaluate other CATE models, including the DR-learner [126] and T-learner [134], with results showing consistent performance. Details are in C.5.

learning models employed have a parameter count comparable to our architecture detailed in Section 3.5.1. In the main text, we consider two types of joint CATE models:[9]

1. **Deep DR-learner**: We develop a traditional deep neural network to predict the doubly robust score *without constructing separate representations for design features and customer covariates*. This benchmark enables us to quantify the benefits of the proposed model architecture compared to traditional deep learning models.

2. **Deep DR-learner with No Design Feature**: This method adopts the previous approach but *utilizes only customer covariates* for the deep neural network. This model helps evaluate the value created solely by learning across experiments, excluding the design features information, in comparison to individual modeling.

Performance metrics

We focus on two performance metrics: predictive accuracy of treatment effects and targeting effectiveness, both evaluated at the offer level. Predictive accuracy is crucial because managers often depend on the predicted treatment effect to evaluate the expected profit from an offer. Targeting effectiveness is equally important, as it informs decisions about who should receive the offer.

Predictive Accuracy. We evaluate the accuracy of a CATE model ($\widehat{\tau}$) in predicting treatment effects for each offer o. To do this, we divide the test set of each offer into $Q = 50$ equally sized groups based on their predicted CATEs. Next, we compute the average treatment effect for each group q_o based on (i) the predicted CATEs, that is, $\widehat{\tau}_{q_o} \equiv \frac{1}{|\#\{i \in q_o\}|} \sum_{i \in q_o} \widehat{\tau}(\mathbf{Z}_o, \mathbf{X}_{o,i})$, and (ii) the actual outcomes observed in the experiments, i.e.,

$$\tau_{q_o} \equiv \frac{1}{|\#\{i : i \in q_o, W_{o,i} = 1\}|} \sum_{i \in q_o, W_{o,i}=1} Y_{o,i} - \frac{1}{|\#\{i \in q_o, W_{o,i} = 0\}|} \sum_{i \in q_o, W_{o,i}=0} Y_{o,i}.$$

Finally, we compute the average root mean squared error (RMSE) across all offers, i.e., $\text{RMSE}(\widehat{\tau}) = \frac{1}{O} \sum_{o=1}^{O} \sqrt{\sum_{q_o=1}^{Q} \left(\widehat{\tau}_{q_o} - \tau_{q_o}\right)^2}$.

[9]Similar to individual modeling, we have estimated various types of CATE models using pooled data. Detailed information on these models and the full set of results can be found in C.5.

Targeting Performance. We assess targeting performance by analyzing each model's ability to distinguish between customers with high treatment effects and those with low treatment effects. Specifically, we choose the Area Under the Targeting Operating Characteristic curve (AUTOC) [210], a common metric that evaluate a model's ability to correctly rank individuals based on their treatment effects. To calculate that value, we first estimate the *Targeting Operating Characteristic* (TOC) for offer o, which is defined as the difference in treatment effects between individuals within the top $\phi \times 100\%$ predicted CATE tier and all individuals. That is,

$$\begin{aligned}\mathrm{TOC}_o(\phi;\widehat{\tau}) = \mathbb{E}\left[Y_{o,i}(W_{o,i}=1) - Y_{o,i}(W_{o,i}=0) | F_{\widehat{\tau}}(\widehat{\tau}(\mathbf{Z}_o,\mathbf{X}_{o,i})) \geqslant 1-\phi\right] - \\ \mathbb{E}\left[Y_{o,i}(W_{o,i}=1) - Y_{o,i}(W_{o,i}=0)\right],\end{aligned}$$

where $F_{\widehat{\tau}}$ is the cumulative distribution function of the predicted CATEs. Then, the AUTOC is defined as $\mathrm{AUTOC}_o(\widehat{\tau}) = \int_0^1 \mathrm{TOC}_o(\phi;\widehat{\tau})d\phi$. Note that a model $\widehat{\tau}$ outperforms another model $\widehat{\tau}'$ in identifying customers in the top $\phi \times 100\%$ CATE group if $\mathrm{TOC}_o(\phi;\widehat{\tau}) > \mathrm{TOC}_o(\phi;\widehat{\tau}')$. Therefore, a higher AUTOC value suggests that the CATE model is more successful at identifying customers who demonstrate the strongest sensitivity toward the intervention, leading to more effective targeting for offer o. Finally, we compute the average AUTOC value across all the offers: $\mathrm{AUTOC}(\widehat{\tau}) = \frac{1}{O}\sum_{o=1}^{O}\mathrm{AUTOC}_o(\widehat{\tau})$.

Results

We implement a bootstrap validation scheme, similar to the one described in [9], to assess the performance of our model. Specifically, we create $B = 20$ pairs of training (70% of the data from each experiment) and test splits (30% of the data from each experiment) for the 274 offers. For each split b, we train a CATE model $\widehat{\tau}^{(b)}$ using the training set and then assess its performance ($\mathrm{RMSE}(\widehat{\tau}^{(b)})$ and $\mathrm{AUTOC}(\widehat{\tau}^{(b)})$) on the corresponding test set. This procedure is replicated B times, allowing us to calculate and report the bootstrap mean and standard deviation for key performance metrics. Table 3.2 presents the bootstrap means and standard deviations of average RMSE and AUTOC values for each approach, across the 274 offers.

Several findings are particularly noteworthy. First, the IRL approach (the first row) achieves the lowest RMSE and the highest AUTOC value, demonstrating its exceptional predictive accuracy and targeting precision. In particular, the IRL model significantly outperforms the joint doubly

Table 3.2: *Predictive Validity for 274 Tested Offers*

Model	RMSE	AUTOC
IRL (10-dim)	7.78 (0.82)	0.77 (0.23)
Joint-DR-DNN	8.29 (1.52)	0.45 (0.20)
Joint-DR-DNN (No Design Features)	8.62 (1.11)	0.48 (0.13)
Individual-DR-DNN	8.57 (1.41)	0.05 (0.20)

Note: We calculated the mean average RMSE and AUTOC values across 274 offers over 20 splits, with standard deviations shown in parentheses. The results demonstrate the effectiveness of models based on deep neural networks. For a complete set of results from other CATE models, please refer to Table C.2 in C.5.

robust DNN model (listed in the second row), suggesting that its superior performance is largely attributable to its network architecture. This supports the argument in Section 3.5.1 that constructing separate representations for offers and customers can enhance predictive accuracy.

Second, all the joint models that leverage information across experiments (rows one to three) outperform the method that estimates the CATE for each offer individually (the final row). Notably, the individual CATE models exhibit a significantly higher RMSE and limited targeting ability (as the mean AUTOC value is close to zero). This highlights the benefits of synthesizing knowledge from multiple experiments.

Third, the deep DR-learner with design features (the second row) performs similarly to the model without them (the third row). This suggests that the traditional deep learning architecture may not effectively capture the generalizable patterns of design features that influence treatment effect heterogeneity.

3.6 Application: Targeting and Tailoring New Promotions

Having demonstrated the predictive validity, we now examine how effectively the IRL model extends its predictions to untested offers and customer segments. We also show how managers can use the IRL model to enhance profitability through tailored offers. Specifically, we show that the model can identify key features and covariates from learned representations, helping managers set optimal levels of these design features for specific customer segments, thus optimizing promotional designs and boosting profitability.

3.6.1 Using IRL for Targeting

We explore two decision scenarios for targeting with our model. First, we assess the model's ability to select customer targets for new, untested offers. This evaluation seeks to determine the model's generalizability across different promotional designs, enabling managers to enhance the effectiveness of new offers without prior testing. Second, we use the model to target existing offers to new, untested customers (i.e., customers who were originally ineligible for such offers). This exercise aims to evaluate the model's adaptability to new customer segments, thereby enabling managers to effectively expand the reach of existing offers.

Targeting for Untested Offers

We start by examining a scenario where the focal company aims to evaluate and make targeting decisions for offers that have not yet been tested. As a case study, we select 55 offers related to personal care products (one of the largest product categories in the data) as the *holdout* offers and use the remaining 219 offers (none of which promoted personal care products) to train the IRL model as well as the other joint-CATE benchmarks.[10] We compare the performance of all joint models against the traditional approach, in which the company conducts experiments for each of the 55 holdout offers and estimates individual CATE models using the actual experimental data.

To implement bootstrap validation, we first divide the actual experimental data for these 55 offers into a 70% training set and a 30% test set for each iteration. We then use the training set to estimate the individual CATE models and evaluate their performance on the test sets. For the IRL and other joint models, we train using data from the 219 offers and evaluate performance on the test set of each of the 55 holdout offers. We conduct 20 splits and report the means and standard deviations of the key performance metrics.

Table 3.3 presents the bootstrap means and standard deviations of average RMSE and AUTOC values for each CATE model across 55 holdout offers. Three key findings emerge. First, the proposed IRL method consistently surpasses other methods in targeting performance, demonstrating

[10]Essentially, we are assessing whether attribute shifts and concept shifts are significant concerns in our empirical application. Regarding concept shifts, the true CATE function for one category may differ substantially from other categories. For attribute shifts, it is important to note that the distribution of design features and eligible customers for the 55 offers varies from the other offers (see C.7.2 for additional details).

superior generalization to untested interventions. Second, the three joint models (rows one to three) significantly outperform the individual CATE models from actual experiment data (last row), highlighting the benefits of leveraging past experiments. Third, while RMSE values are similar across models using past experiments, the deep learning model integrating design features and customer covariates (second row) shows poorer targeting performance (smaller AUTOC) compared to the model excluding design features (third row). This confirms the advantage of separating design and customer representations for better generalization.

Table 3.3: *Targeting Performance for Offers Related to Personal Care Products*

Model	RMSE	AUTOC
IRL (10-dim)	3.28 (0.56)	0.67 (0.15)
Joint-DR-DNN	3.32 (0.56)	0.31 (0.10)
Joint-DR-DNN (No Design Features)	3.62 (0.39)	0.54 (0.11)
Actual (Individual-DR-DNN)	5.44 (0.60)	0.02 (0.14)

Note: We calculated the mean average RMSE and AUTOC values across 55 holdout offers, with standard deviations shown in parentheses. The results demonstrate the effectiveness of models based on deep neural networks. For a complete set of results from other CATE models, please refer to Table C.3 in C.5.

Targeting for Untested Customers

Next, we investigate a scenario in which the focal company aims to predict treatment effects and make targeting decisions for customer segments who were previously ineligible for certain offers. As a case study, we focus on "customers over the age of 40 and offers in the wine category," and consider a scenario where the company had not previously targeted this demographic for their wine-related promotions. Our IRL model uses data from past wine promotions targeted at younger customers, combined with offers from other product categories that were tested across various age groups, to predict the treatment effects of wine promotions on this older customer segment.

Specifically, we use experimental samples from wine offers involving customers over the age of 40 (from both treatment and control groups) as our *holdout* data. This dataset includes 31 distinct offers, comprising a total of 783,671 observations. Similar to Section 3.6.1, we apply the bootstrap validation approach with 20 splits for performance evaluation. We first divide the holdout data into a 70% training set and a 30% test set for each bootstrap split. Then, the training set is used to

develop the individual CATE models, whose performance is then assessed on the test sets. For IRL and other joint models, we use all the non-holdout data for training and then assess their performance on the test set derived from the holdout data.

Table 3.4 shows the performance metrics of various models on the holdout data. First, the IRL model (the first row) achieves the best results, while the joint model using a traditional deep learning architecture (the second row) displays comparable performance. Second, models that leverage data across experiments (rows one to three) outperform the individual CATE models for each offer (the final row). Notably, the model that excludes design features is the least effective among the joint models but still significantly outperforms the individual CATE models. These findings suggest that the IRL model and (other joint models) can enhance targeting performance on new target populations from previously tested offers by synergizing past experiments.

Table 3.4: *Targeting Performance for Wine Offers on Customers with Age over 40*

Model	RMSE	AUTOC
IRL (10-dim)	5.58 (0.77)	1.27 (0.36)
Joint-DR-DNN	5.75 (0.67)	1.16 (0.37)
Joint-DR-DNN (No Design Features)	6.27 (0.83)	0.98 (0.39)
Actual (Individual-DR-DNN)	6.65 (0.98)	0.15 (0.45)

Note: We calculated the mean average RMSE and AUTOC values using holdout data, which includes 783,671 observations from 31 distinct offers, with standard deviations shown in parentheses. The results demonstrate the effectiveness of models based on deep neural networks. For a complete set of results from other CATE models, please refer to Table C.4 in C.5.

3.6.2 Using IRL for Tailoring

Typically, managers tailor their promotions through a two-step process: first, they create customer segments based on observed covariates, such as category preferences or buying behaviors. Then, they create targeted offers for each segment based on their objectives (e.g., customers who usually make smaller purchases should receive greater discounts to incentivize them to buy more) or prior beliefs (e.g., offers with stricter dollar requirements usually work for customers who have made larger purchases). In this section, we demonstrate how the IRL model can provide a data-driven approach to this decision-making process, allowing companies to tailor promotional offers more effectively to different customer segments and enhance profitability.

Insights from Incrementality Representations

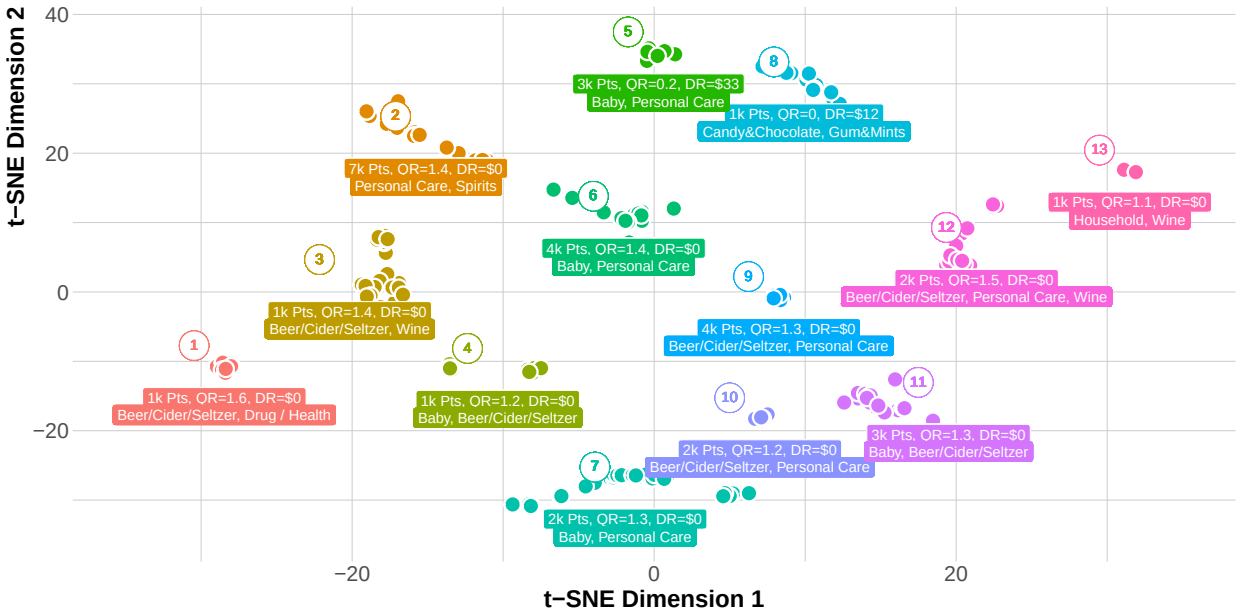
A key feature of the IRL model is its ability to generate low-dimensional *incrementality representations* for both offer features (ζ_o) and customer covariates ($\chi_{o,i}$). We illustrate how managers can use these representations to identify key decision factors for optimal promotion design. Specifically, we conduct segmentation analyses of the incrementality representations, providing insights that can further inform managers' tailoring decisions. (The results discussed below are derived from one bootstrap split detailed in Section 3.5.2).

Offer Design Features. By design, offers with similar incrementality representations exhibit comparable response patterns from the same customer group (as the model generates similar treatment effects for these offers when applied to the same customer). Consequently, managers can utilize these representations to (i) identify which previously tested offers are effective for the same customer and (ii) discern the key design features influencing these patterns. With this knowledge, managers can concentrate on the design features that significantly impact incrementality representations, and therefore treatment effect heterogeneity, enabling them to tailor promotions more effectively.

To achieve this, we extract the learned representation for the design of the 274 existing offers (denoted as ζ_o following the notation in Figure 3.1). We then employ t-distributed stochastic neighbor embedding (t-SNE) [197] to project these representations into a two-dimensional space and use the DBSCAN algorithm [69] to identify "offer clusters" with similar incrementality representations. Each cluster is profiled based on the design features ($\mathbf{Z}_o$) of the offers it contains. For continuous features (e.g., points awarded), we calculate the cluster average; for categorical features (e.g., product category), we identify the most frequently occurring labels within the cluster.

Figure 3.4 displays the t-SNE map and the clusters generated using incrementality representations from the proposed IRL model. Each point on the map represents a promotional offer, color-coded according to the clusters identified by the DBSCAN algorithm. Labels with colored backgrounds provide key information for each cluster, including average reward points, required quantities, required dollar amounts, and the top two promoted product categories.

Figure 3.4 suggest that clusters based on incrementality representations provide a nuanced perspective on offers with similar patterns of heterogeneous treatment effects. For instance, consider the clusters related to baby and personal care products. For these categories, the incrementality representations identify three distinct clusters (clusters 5, 6, and 7 in Figure 3.4). Cluster 7 is characterized by lower reward points (2,000 points) and quantity requirement (QR = 1.3), while Clusters 5 and 6 feature higher reward points (3,000 points and 4,000 points, respectively), differing mainly in their reward claim requirements. Specifically, Cluster 6 demands a higher quantity requirement (QR = 1.4), whereas Cluster 5 requires purchases to meet a certain dollar amount (DR = $33).

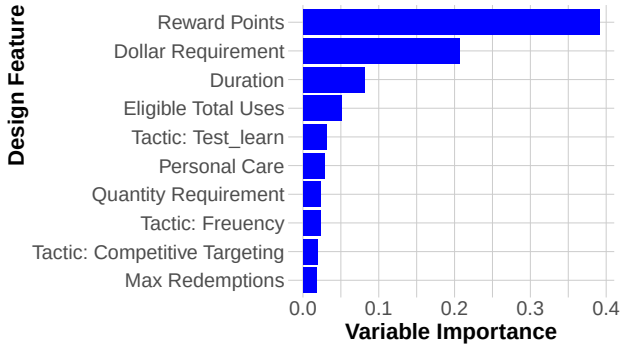


Note. Each point on the map corresponds to a promotional offer, with colors indicating the clusters determined by the DBSCAN algorithm. Labels with colored backgrounds show the average reward points (Pts, set as zero for offers requiring minimum dollar spending), required quantities (QR), required dollar amounts (DR, set as zero for offers requiring certain product quantities), and the top two promoted categories for each cluster.

Figure 3.4: *t-SNE Maps for Learned Representations (ζ_o)*

To further understand what are driving the differences behind the incrementality representations, we conduct a statistical analysis to identify the features determining cluster membership. Specifically, we use XGBoost to construct a multi-class classification model that predicts the cluster membership of each offer based on its design features. Then, the ten features with the highest variable importance are shown in Figure 3.5. The classification models show that clusters

of incrementality representations are influenced by a diverse set of factors, with reward points accounting for 40% of the importance, requirements making up 20%, and additional features such as duration, product categories, and promotional tactics collectively contributing 35%.



Note: Variable importance is measured by the average gain in accuracy across all splits where the variable is utilized.

Figure 3.5: *Variable Importance of Top 10 Design Features for Cluster Classification Models*

For comparison, we also conduct a similar analysis with the original design features ($\mathbf{Z}_o$) and present the results in C.7.3. We find that clustering offers using raw features provides less discriminative power in separating offers (Figure C.4). In partciular, clusters based on raw design features (Figure C.5) are overwhelmingly influenced by just two features: reward points, which constitute 80% of the classification outcomes, and dollar requirements, which account for 17%.

Customer Covariates. We conduct a similar analysis comparing clusters from the learned representations of customer covariates ($\chi_{o,i}$) to those derived directly from (rescaled) customer covariates ($\mathbf{X}'_{o,i}$).[11] Given the significantly larger data size (millions of customers), we opt for K-means clustering over density-based methods to enhance computational efficiency. In this analysis, we choose three distinct customer segments for illustrative purposes. We perform this analysis using both (a) the incrementality representations from the test data covariate samples[12] and (b) the rescaled covariate values ($\mathbf{X}'_{o,i}$). Once the segments are established, we profile each segment by averaging the customers' past activities. We summarize both "general" behavior,

[11]To prevent variables with larger scales from dominating the clustering results, we standardize all continuous variables using the Z-score transformation.

[12]We use the test data for this analysis as the subsequent section validates the treatment effect heterogeneity across these derived customer segments using the test set.

which includes all receipts uploaded by the customer in the 90 days prior to receiving the offer, and "offer-related" behavior, which pertains specifically to receipts that included products from each focal offer. Table 3.5 presents both sets of results, with segments sorted by the average total number of receipts (first column).

Table 3.5: *Summary of Covariate Segments Based on (a) Incrementality Representations ($\chi_{o,i}$) and (b) Rescaled Covariated Values ($\mathbf{X}'_{o,i}$)*

(a) Segments Based on Incrementality Representations ($\chi_{o,i}$)

General		Offer-related		Proportion
Receipts	AOV	Receipts	AOV	
18.8	$ 38	0.34	$ 3.8	55%
73.8	$ 44	0.36	$ 2.2	40%
225.7	$ 125	0.86	$ 5.2	5%

(b) Segments Based on Scaled Covariate Values ($X'_{o,i}$)

General		Offer-related		Proportion
Receipts	AOV	Receipts	AOV	
23.1	$ 37	0.31	$ 2.6	23%
51.3	$ 51	0.39	$ 3.6	49%
71.0	$ 48	0.40	$ 3.0	28%

Upon initial examination, both approaches appear to categorize customers based on their engagement level with the platform, as indicated by the disparities in the number of receipts shown in the first column of Tables 3.5a and 3.5b). However, segmentation via incrementality representations uncovers more nuanced distinctions. Notably, Table3.5a shows a broader spread in the total number of receipts compared to Table 3.5b, suggesting that incrementality representations offer a stronger discriminative power. Furthermore, the incrementality representations distinguish customers by average order value (AOV), as detailed in the second column of Tables 3.5a. In contrast, segmentation based on rescaled covariate values fails to differentiate customers by AOV, as illustrated in the second column of Table 3.5b. This distinction is important, given that customers with higher AOVs are more likely to meet quantity or spending requirements, making AOV a critical factor in influencing treatment effects (as illustrated in Section 3.6.2 below).

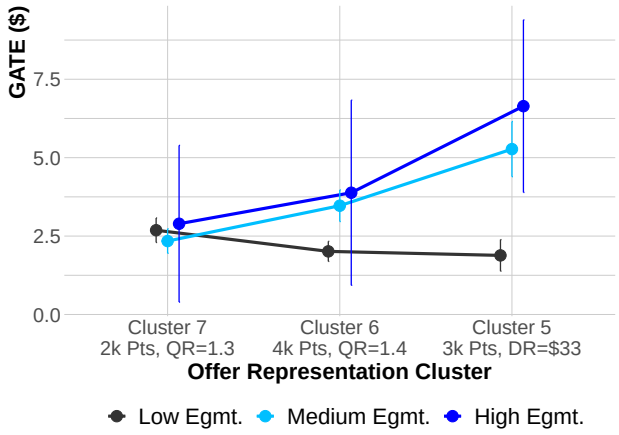
Moreover, segments derived from incrementality representations display distinct variations in offer-related behaviors, as detailed in columns 3 and 4 of Table 3.5a. This contrasts sharply to segments based on original covariate values, which show no significant differences in uploaded receipts for promoted items, as observed in columns 3 and 4 of Table 3.5b.

Assessing Treatment Effect Heterogeneity

So far, our cluster analysis has shown that incrementality representations capture more nuanced variations in design features and customer covariates. We now focus on illustrating *how*

treatment effect heterogeneity varies across the offer clusters and customer segments identified by the incrementality representations. To achieve this, we label each observation in the *test data* with its corresponding offer cluster from Figure 3.4 and engagement segment from Table 3.5a. We then calculate the group average treatment effect (GATE) for each combination of offer cluster and engagement segment, using the actual outcomes from the test data.

Figure 3.6 shows the GATEs for three clusters associated with "baby and personal care" products across three engagement segments, illustrating distinct treatment effect patterns among different segments. First, the medium and high engagement segments generally display higher incremental spending than the low engagement segment, except in scenarios with relatively low reward points (Cluster 7), where the GATEs show no significant differences across engagement levels. Second, increasing the reward points or requirements leads to significant differences among customer segments. For the medium and high engagement segments, higher points or requirements (Clusters 5 and 6) enhance the treatment effect, whereas they result in less effective offers for low engagement customers. Finally, dollar requirements prove more effective than quantity requirements for medium and high engagement segments, as demonstrated by higher treatment effects for offers in Cluster 5 compared to those in Cluster 6.



Note: Each point reports the group average treatment effect together with two-standard-error interval. The offer representation clusters align with those outlined in Figure 3.4, and the engagement segments correspond to those outlined in Table3.5a.

Figure 3.6: *Treatment Effects by Offer Clusters and Customer Segments Derived from Incrementality Representations* $(\zeta_o, \chi_{o,i})$

In summary, the analysis demonstrates that clustering offers and customers based on their incrementality representations from the IRL model uncovers meaningful patterns related to design

features, customer behaviors, and treatment effects, which are not as evident when using the original features and covariates. By segmenting samples based on the key variables identified through these incrementality representations, companies can more effectively detect variations in treatment effects and identify key design features and customer segments driving such variations.[13]

Tailoring Promotions Using Segment Incrementality Plots

Building on our analysis, we identified two critical design features — reward points and dollar requirements — that influence treatment effect heterogeneity across three customer segments. In this section, we develop an interpretable machine learning tool for the IRL model, enabling managers to tailor promotions to maximize predicted profitability rather than relying on business intuition.

Suppose a manager is considering a specific offer, like a "30-day, one-time-use baby product offer that can be combined with other offers," but is unsure about the optimal reward points and required dollar amount due to varied customer responses. We recommend using a post-hoc model interpretation method. By leveraging predictions from the IRL model, this approach can determine the optimal reward points and dollar requirements for different customer segments, tailoring the offer to maximize effectiveness.

Specifically, we introduce a model visualization framework named the *Segment Incrementality Plot* (SIP). This plot is created by calculating the predicted treatment effects for different customer segments using the IRL model, considering variations in reward points and dollar requirements for the offer. We then calculate the predicted incremental profits across segments based on these treatment effects. The construction of the SIP involves the following steps:

1. Randomly sample covariate observations from pre-defined customer segments for computational efficiency. Here, we use the low, medium, and high engagement segments identified in Table 3.5a for illustration (denoted as $\mathcal{S}_{\text{Low}}$, $\mathcal{S}_{\text{Medium}}$, and $\mathcal{S}_{\text{High}}$).

2. Create hypothetical offers with a range of reward points $\{Z_{\text{new}}^{\text{point}}\}$ and dollar requirements $\{Z_{\text{new}}^{\text{dollar}}\}$, while keeping all other design features constant ($\overline{\mathbf{Z}}_{\text{new}}$).

[13]For comparison, an analysis for clusters based on original design features and scaled customer covariates is detailed in C.7.4. The results show minimal significant differences across defined clusters and segments. In contrast, as depicted in Figure 3.6, incrementality representations more effectively elucidate treatment effect heterogeneity.

3. Compute predicted treatment effects for every combination of $Z_{\text{new}}^{\text{point}}$, $Z_{\text{new}}^{\text{dollar}}$ and $\mathbf{X}_{o,i}$ within $\mathcal{S}_{\text{Low}}$, $\mathcal{S}_{\text{Medium}}$, $\mathcal{S}_{\text{High}}$, that is, $\widehat{\tau}\left(\widehat{\mathbf{\Phi}}(Z_{\text{new}}^{\text{point}}, Z_{\text{new}}^{\text{dollar}}, \overline{\mathbf{Z}}_{\text{new}}), \widehat{\mathbf{\Psi}}(\mathbf{X}_{o,i})\right)$.

4. Calculate the average predicted treatment effect for each combination of $Z_{\text{new}}^{\text{point}}$, $Z_{\text{new}}^{\text{point}}$ and each segment (i.e., the segment incrementality):

$$\text{SI}(Z_{\text{new}}^{\text{point}}, Z_{\text{new}}^{\text{dollar}}, \mathcal{S}_{(\cdot)}) = \frac{1}{|\mathcal{S}_{(\cdot)}|} \sum_{i \in \mathcal{S}_{(\cdot)}} \widehat{\tau}\left(\widehat{\mathbf{\Phi}}(Z_{\text{new}}^{\text{point}}, Z_{\text{new}}^{\text{dollar}}, \overline{\mathbf{Z}}_{\text{new}}), \widehat{\mathbf{\Psi}}(\mathbf{X}_{o,i})\right).$$

5. Calculate the predicted incremental profits of the new offer on each customer using $\text{SI}(Z_{\text{new}}^{\text{point}}, Z_{\text{new}}^{\text{dollar}}, \mathcal{S}_{(\cdot)})$. [14] Then, visualize $\text{Profit}(Z_{\text{new}}^{\text{point}}, Z_{\text{new}}^{\text{dollar}}, \mathcal{S}_{(\cdot)})$ for each segment.

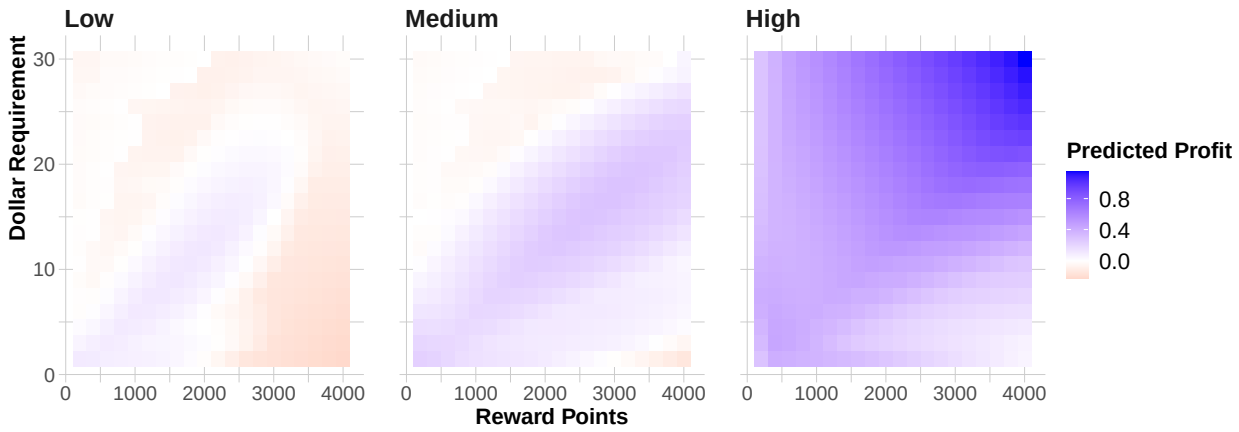
The SIP offers significant advantages as a decision support tool. First, unlike partial dependence plots [77], which average effects across all customers, the SIP allows for the identification of differential interactions between customer engagement levels and offer design features. Second, in contrast to individual conditional expectation plots [84], which focus on predictions for each individual, the SIP provides insights at the segment level. This facilitates the interpretation of results by illustrating how different segments respond to various promotional strategies. Third, the SIP works for any predefined customer segments. Therefore, managers can apply the SIP to segments from any method and evaluate whether treatment effect heterogeneity exists for those segments based on the SIP.

We now illustrate how to apply the SIP to tailor promotional offers for customer segments identified using incrementality representations. Figure 3.7 shows the SIP of profitability for a baby product offer, generated using the IRL model, where reward points range from 200 to 4000 and the dollar requirement ranges from \$1.5 to \$30. The blue zones in the plots indicate profitable regions, which differ across customer segments. (We also present the SIP of the predicted treatment effects in C.7.5.)

The SIP suggests that for the high-engagement segment (right-most figure), higher dollar

[14] Since profitability data were not directly available from the focal company, we employ the following formula as a proxy for profit: $\text{Profit}(Z_{\text{new}}^{\text{point}}, Z_{\text{new}}^{\text{dollar}}, \mathcal{S}_{(\cdot)}) = \text{SI}(Z_{\text{new}}^{\text{point}}, Z_{\text{new}}^{\text{dollar}}, \mathcal{S}_{(\cdot)}) \times \text{Avg. Rev.} - Z_{\text{new}}^{\text{point}} \times \text{Cost per Point} \times \text{Reward Claim Rate}(\mathcal{S}_{(\cdot)})$, where Avg. Rev. represents the average revenue from one dollar of incremental purchase, Cost per Point indicates the actual dollar cost associated with each reward point, and $\text{Reward Claim Rate}(\mathcal{S}_{(\cdot)})$ reflects the segment-specific reward claim rates, which are derived from previously observed reward claim behaviors across each engagement segment.

requirements significantly increases the predicted incremental profit. These customers are more tolerant of higher spending thresholds, making offers with substantial rewards more profitable. In contrast, designing promotions for low- and medium-engagement customers requires a nuanced approach. Offers with substantial rewards and small spending requirements are often unprofitable for these groups, as the additional spending does not offset the promotional costs. Additionally, moderate rewards with high spending requirements tend to attract only low-value "cherry pickers," further diminishing profits in these segments.



Note: We set the reward points within the range of $\{200, \cdots, 4000\}$ and the dollar requirement within $\{\$1.5, \cdots, \$30\}$.

Figure 3.7: *Segment Incrementality Plot of Profitability for the Focal Offer Design*

Combining the results from the three figures, it becomes clear that optimal offer designs vary across segments: For the low-engagement segment, an offer of 1,600 points with a minimum $9 purchase yields a modest profit of $0.11 per customer. In the medium-engagement segment, 3,000 points with a minimum $16.5 purchase generates $0.30 profit per customer, while in the high-engagement segment, 4,000 points with a minimum $30 purchase results in $1.15 profit per customer. It is crucial to note that applying the strategy designed for the high-engagement segment to the lower segments would not only fail to generate profit but would actually incur losses of about $0.02 per customer. This underscores the importance of tailoring promotional offers to distinct customer segments to achieve optimal profit outcomes.

In conclusion, the discriminative power of the incrementality representations offers a nuanced understanding of how to optimize promotional personalization for diverse customer populations. These insights not only help understand treatment effect heterogeneity but also boost profitability

by navigating the complexities of promotional design. By leveraging these decision-support tools, marketers can implement tailored strategies that ensure each customer segment receives an optimized offer, effectively aligning marketing efforts with individual preferences and behaviors.

3.7 Conclusion and Future Directions

In pursuit of personalization, companies aim to target and tailor marketing interventions to maximize incremental value. Traditional experimental approaches are limited by the capacity to test only a few interventions within specific segments. To overcome this, we introduce the IRL framework. This approach leverages data from past experiments to target and tailor personalized interventions more effectively. By extracting low-dimensional representations that capture generalizable patterns in treatment effect heterogeneity, the IRL framework accurately predicts CATEs for tested interventions and generalizes to new, untested interventions and customer segments. This method helps companies address key personalization challenges, such as tailoring interventions for specific segments to maximize impact and identifying responsive segments for new interventions.

We demonstrate the superior performance of the IRL model over traditional methods through its application to CPG promotional campaigns. By synergizing past experiments, the IRL method not only enhances the accuracy of treatment effect predictions and targeting effectiveness for previously tested interventions but also adeptly extends to untested interventions or segments. Essentially, the IRL model addresses two major challenges in personalizing interventions: managing the complexity of a high-dimensional decision space and overcoming the cold-start problem for untested interventions. Through this empirical application, we also show how firms can use the learned representations to identify critical design features and customer characteristics that significantly influence the effectiveness of promotions. Furthermore, we introduce a model interpretation tool that assists companies in designing optimally tailored promotions for various customer segments, further enhancing the practical utility of the IRL framework.

While our research provides a valuable tool for researchers and managers to design personalized interventions, it also presents several limitations that open avenues for future research. First, a practical concern identified in Theorem 3.2 is managing both *attribute shift* and *concept shift*,

which can undermine the accuracy of our IRL model when applied to untested interventions or customer segments that deviate from historical data. To mitigate these issues, one can apply active learning techniques [e.g., 119, 192] and target new data collection efforts toward offers or customer segments with high uncertainty in their CATE predictions, thereby enhancing the model's generalization capabilities. Furthermore, model fine-tuning strategies in transfer learning [e.g., 5, 182] could be employed to refine predictions based on newly collected, small-scale experiments, ensuring the model's continued accuracy in dynamic environments.

Second, beyond improving generalizability through new data collection, future research could also investigate optimal strategies for determining which interventions should and should not be learned together, and what information should be transferred between them. Negative transfer can occur when interventions that elicit markedly different responses from the same customers are combined in learning processes [203]; in such cases, learning from individual experiments may outperform learning across multiple experiments. The existing literature [e.g., 215, 118] provides potential strategies to mitigate this issue by specifying (a) which parameters should be shared across experiments, and (b) which features or covariates should be used for experiment, and to what extent this should be transferred. Integrating these approaches into our IRL model could further improve its accuracy and targeting effectiveness.

Third, while we provide a flexible tool for managers to interpret model behaviors, it currently supports only post-hoc exploration of model predictions at the segment level. Future research could investigate the application of various explainable machine learning tools to enable managers to better understand treatment effect heterogeneity. For instance, one could assess how different design features and customer covariates most significantly contribute to the treatment effect using frameworks such as feature attribution [e.g., 167, 142] or counterfactual explanation [e.g., 160, 4], thereby providing deeper insights into the behaviors of the IRL model.

Finally, while we have shown the benefits of synergizing experiments in a promotional setting, exploring this approach's applicability across various marketing applications could be highly valuable. For instance, an e-commerce company could use insights from email or push notification experiments to gauge customer responsiveness to promotional offers. Additionally, responses to different product recommendation algorithms could help better predict the treatment effects of various email communications and tailor them more effectively. Investigating the effectiveness

of our framework in diverse contexts could reveal broader implications for when and how personalized marketing strategies can significantly benefit from integrated experimental insights.

We hope that these limitations inspire further research on synergizing experiments and incrementality representation learning.

Bibliography

[1] Andrew B Abel. Classical measurement error with several regressors. *Wharton School of the*, 2018.

[2] Alessandro Acquisti and Ralph Gross. Predicting social security numbers from public data. *Proceedings of the National Academy of Sciences*, 106(27):10975–10980, 2009.

[3] Anish Agarwal and Rahul Singh. Causal inference with corrupted data: Measurement error, missing values, discretization, and differential privacy. *arXiv preprint arXiv:2107.02780*, 2021.

[4] Emanuele Albini, Jason Long, Danial Dervovic, and Daniele Magazzeni. Counterfactual shapley additive explanations. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1054–1070, 2022.

[5] Tagrid Alshalali and Darsana Josyula. Fine-tuning of pre-trained deep learning models with extreme learning machine. In *International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 469–473. IEEE, 2018.

[6] Kareem Amin, Alex Kulesza, Andres Munoz, and Sergei Vassilvtiskii. Bounding user contributions: A bias-variance trade-off in differential privacy. In *International Conference on Machine Learning*, pages 263–271. PMLR, 2019.

[7] Asim Ansari and Carl F Mela. E-customization. *Journal of Marketing Research*, 40(2):131–145, 2003.

[8] Neeraj Arora and Ty Henderson. Embedded premium promotion: Why it works and how to make it more effective. *Marketing Science*, 26(4):514–531, 2007.

[9] Eva Ascarza. Retention futility: Targeting high-risk customers might be ineffective. *Journal of Marketing Research*, 55(1):80–98, 2018.

[10] Eva Ascarza, Scott A Neslin, Oded Netzer, Zachery Anderson, Peter S Fader, Sunil Gupta, Bruce GS Hardie, Aurélie Lemmens, Barak Libai, David Neal, et al. In pursuit of enhanced customer retention management: Review, key issues, and future directions. *Customer Needs and Solutions*, 5:65–81, 2018.

[11] Eva Ascarza, Oded Netzer, and Bruce GS Hardie. Some customers would rather leave without saying goodbye. *Marketing Science*, 37(1):54–77, 2018.

[12] Susan Athey. Beyond prediction: Using big data for policy problems. *Science*, 355(6324):483–485, 2017.

[13] Susan Athey, Raj Chetty, Guido W Imbens, and Hyunseung Kang. The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Technical report, National Bureau of Economic Research, 2019.

[14] Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

[15] Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

[16] Susan Athey and Stefan Wager. Estimating treatment effects with causal forests: An application. *Observational Studies*, 5(2):37–51, 2019.

[17] Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021.

[18] Philipp Bach, Victor Chernozhukov, Malte S Kurz, and Martin Spindler. Doubleml-an object-oriented implementation of double machine learning in python. *Journal of Machine Learning Research*, 23(53):1–6, 2022.

[19] Patrick Bachmann, Markus Meierer, and Jeffrey Näf. The role of time-varying contextual factors in latent attrition models for customer base analysis. *Marketing Science*, 40(4):783–809, 2021.

[20] Randall Balestriero, Jerome Pesenti, and Yann LeCun. Learning in high dimension always amounts to extrapolation. *arXiv preprint arXiv:2110.09485*, 2021.

[21] Imre Bárány and Zoltán Füredi. On the shape of the convex hull of random points. *Probability Theory and Related Fields*, 77:231–240, 1988.

[22] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[23] Hamsa Bastani. Predicting with proxies: Transfer learning in high dimension. *Management Science*, 67(5):2964–2984, 2021.

[24] Erich Battistin and Andrew Chesher. Treatment effect estimation with covariate measurement error. *Journal of Econometrics*, 178(2):707–715, 2014.

[25] Kapil Bawa. Modeling inertia and variety seeking tendencies in brand choice behavior. *Marketing Science*, 9(3):263–278, 1990.

[26] Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.

[27] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

[28] Shai Ben-David and Reba Schuller. Exploiting task relatedness for multiple task learning. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings*, pages 567–580. Springer, 2003.

[29] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[30] Laurent Bonnasse-Gahot. Interpolation, extrapolation, and local generalization in common neural networks. *arXiv preprint arXiv:2207.08648*, 2022.

[31] Bryan W Brown. The identification problem in systems nonlinear in the variables. *Econometrica: Journal of the Econometric Society*, pages 175–196, 1983.

[32] Maya Burhanpurkar, Zhun Deng, Cynthia Dwork, and Linjun Zhang. Scaffolding sets. *arXiv preprint arXiv:2111.03135*, 2021.

[33] Raymond J Carroll and Peter Hall. Low order approximations in deconvolution and regression with errors in variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 66(1):31–46, 2004.

[34] Raymond J Carroll, Kathryn Roeder, and Larry Wasserman. Flexible parametric measurement error models. *Biometrics*, 55(1):44–54, 1999.

[35] Fanglin Chen, Xiao Liu, Davide Proserpio, and Isamar Troncoso. Product2vec: Leveraging representation learning to model consumer product choice in large assortments. *NYU Stern School of Business*, 2022.

[36] Huigang Chen, Totte Harinen, Jeong-Yoon Lee, Mike Yung, and Zhenyu Zhao. CausalML: Python package for causal machine learning, 2020.

[37] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[38] Xiaohong Chen, Han Hong, and Elie Tamer. Measurement error models with auxiliary data. *The Review of Economic Studies*, 72(2):343–366, 2005.

[39] Yi-Wen Chen, Eva Ascarza, and Oded Netzer. Policy-aware experimentation: Strategic sampling for optimized targeting policies. *Columbia Business School Research Paper*, (5044549), 2024.

[40] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018.

[41] Victor Chernozhukov, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val. Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Technical report, National Bureau of Economic Research, 2018.

[42] Victor Chernozhukov, Guido W Imbens, and Whitney K Newey. Instrumental variable estimation of nonseparable models. *Journal of Econometrics*, 139(1):4–14, 2007.

[43] Andrew Chesher. The effect of measurement error. *Biometrika*, 78(3):451–462, 1991.

[44] Andrew Chesher. Identification in nonseparable models. *Econometrica*, 71(5):1405–1441, 2003.

[45] Pradeep K Chintagunta, Liqiang Huang, Wei Miao, and Wanqing Zhang. Measuring seller response to buyer-initiated disintermediation: Evidence from a field experiment on a service platform. *Available at SSRN 4423917*, 2023.

[46] François Chollet et al. Keras. https://keras.io, 2015.

[47] Aloni Cohen. Attacks on deidentification's defenses. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1469–1486, 2022.

[48] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

[49] Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.

[50] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

[51] Øystein Daljord, Carl F Mela, Jason MT Roos, Jim Sprigg, and Song Yao. The design and targeting of compliance promotions. *Marketing Science*, 42(5):866–891, 2023.

[52] Irina Degtiar and Sherri Rose. A review of generalizability and transportability. *Annual Review of Statistics and Its Application*, 10:501–524, 2023.

[53] Alex Deng, Ya Xu, Ron Kohavi, and Toby Walker. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, pages 123–132, 2013.

[54] Floris Devriendt, Darie Moldovan, and Wouter Verbeke. A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics. *Big Data*, 6(1):13–41, 2018.

[55] Ryan Dew, Asim Ansari, and Olivier Toubia. Letting logos speak: Leveraging multiview representation learning for data-driven branding and logo design. *Marketing Science*, 41(2):401–425, 2022.

[56] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. *Advances in Neural Information Processing Systems*, 30, 2017.

[57] Tony Duan, Avati Anand, Daisy Yi Ding, Khanh K Thai, Sanjay Basu, Andrew Ng, and Alejandro Schuler. Ngboost: Natural gradient boosting for probabilistic prediction. In *International conference on machine learning*, pages 2690–2700. PMLR, 2020.

[58] Jean-Pierre Dubé, Zheng Fang, Nathan Fong, and Xueming Luo. Competitive price targeting with smartphone coupons. *Marketing Science*, 36(6):944–975, 2017.

[59] Jean-Pierre Dubé, Günter J Hitsch, and Peter E Rossi. State dependence and alternative explanations for consumer inertia. *The RAND Journal of Economics*, 41(3):417–445, 2010.

[60] Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*, pages 1–12. Springer, 2006.

[61] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer, 2006.

[62] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[63] Paul B Ellickson, Wreetabrata Kar, and James C Reeder. Estimating marketing component effects: Double machine learning from targeted digital promotions. *Marketing Science*, 0(0), 2022.

[64] Paul B Ellickson, Wreetabrata Kar, and James C Reeder III. Estimating marketing component effects: Double machine learning from targeted digital promotions. *Marketing Science*, 0(0):1–22, 2023.

[65] Paul B Ellickson, Wreetabrata Kar, James C Reeder III, and Guang Zeng. Using contextual embeddings to predict the effectiveness of novel heterogeneous treatments. *Available at SSRN 4845956*, 2024.

[66] Tülin Erdem and Michael P Keane. Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets. *Marketing Science*, 15(1):1–20, 1996.

[67] Timothy Erickson and Toni M Whited. Two-step gmm estimation of the errors-in-variables model using high-order moments. *Econometric Theory*, 18(3):776–799, 2002.

[68] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pages 1054–1067, 2014.

[69] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

[70] European Data Protection Board. Guidelines 01/2025 on pseudonymisation, 2025. Accessed: 2025-02-04.

[71] Peter S Fader and Bruce GS Hardie. Incorporating time-invariant covariates into the Pareto/NBD and BG/NBD models. *Technical Note*, 2007.

[72] Peter S Fader and Bruce GS Hardie. Customer-base valuation in a contractual setting: The perils of ignoring heterogeneity. *Marketing Science*, 29(1):85–93, 2010.

[73] Peter S Fader, Bruce GS Hardie, and Ka Lok Lee. "Counting your customers" the easy way: An alternative to the Pareto/NBD model. *Marketing Science*, 24(2):275–284, 2005.

[74] Peter S Fader, Bruce GS Hardie, and Jen Shang. Customer-base analysis in a discrete-time noncontractual setting. *Marketing Science*, 29(6):1086–1108, 2010.

[75] Peter S Fader and James M Lattin. Accounting for heterogeneity and nonstationarity in a cross-sectional model of consumer purchase behavior. *Marketing Science*, 12(3):304–317, 1993.

[76] Federal Trade Commission. No, hashing still doesn't make your data anonymous, July 2024. Accessed: 2025-02-04.

[77] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

[78] Jerome H Friedman. Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378, 2002.

[79] Robb Fry and S McManus. Smooth bump functions and the geometry of banach spaces: a brief survey. *Expositiones Mathematicae*, 20(2):143–183, 2002.

[80] Sebastian Gabel, Daniel Guhl, and Daniel Klapper. P2v-map: Mapping market structures for large retail assortments. *Journal of Marketing Research*, 56(4):557–580, 2019.

[81] Sebastian Gabel and Artem Timoshenko. Product choice with large assortments: A scalable deep-learning model. *Management Science*, 68(3):1808–1827, 2022.

[82] Badih Ghazi, Charlie Harrison, Arpana Hosabettu, Pritish Kamath, Alexander Knop, Ravi Kumar, Ethan Leeman, Pasin Manurangsi, and Vikas Sahu. On the differential privacy and interactivity of privacy sandbox reports. *arXiv preprint arXiv:2412.16916*, 2024.

[83] Ira Globus-Harris, Declan Harrison, Michael Kearns, Aaron Roth, and Jessica Sorrell. Multicalibration as boosting for regression. *arXiv preprint arXiv:2301.13767*, 2023.

[84] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 2015.

[85] Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *International Conference on Learning Theory*, pages 297–299. PMLR, 2018.

[86] Negin Golrezaei, Hamid Nazerzadeh, and Paat Rusmevichientong. Real-time optimization of personalized assortments. *Management Science*, 60(6):1532–1551, 2014.

[87] Fusun Gonul and Kannan Srinivasan. Modeling multiple sources of heterogeneity in multinomial logit models: Methodological and managerial issues. *Marketing Science*, pages 213–229, 1993.

[88] Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. *arXiv preprint arXiv:2109.05389*, 2021.

[89] Brett R Gordon, Robert Moakler, and Florian Zettelmeyer. Predictive incrementality by experimentation (pie) for ad measurement. *arXiv preprint arXiv:2304.06828*, 2023.

[90] Justin Grimmer, Solomon Messing, and Sean J Westwood. Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Analysis*, 25(4):413–434, 2017.

[91] Peter M Guadagni and John DC Little. A logit model of brand choice calibrated on scanner data. *Marketing Science*, 2(3):203–238, 1983.

[92] Robin Marco Gubela, Stefan Lessmann, Johannes Haupt, Annika Baumann, Tillmann Radmer, and Fabian Gebert. Revenue uplift modeling. *Machine Learning for Marketing Decision Support*, 2017.

[93] Leo Guelman, Montserrat Guillén, and Ana M Pérez-Marín. Random forests for uplift modeling: an insurance customer retention case. In *International Conference on Modeling and Simulation in Engineering, Economics and Management*, pages 123–133. Springer, 2012.

[94] Leo Guelman, Montserrat Guillén, and Ana M Pérez-Marín. Uplift random forests. *Cybernetics and Systems*, 46(3-4):230–248, 2015.

[95] Yongyi Guo, Dominic Coey, Mikael Konutgan, Wenting Li, Chris Schoener, and Matt Goldman. Machine learning for variance reduction in online experiments. *Advances in Neural Information Processing Systems*, 34, 2021.

[96] Larry Han, Xuan Wang, and Tianxi Cai. On the evaluation of surrogate markers in real world data settings. *arXiv preprint arXiv:2104.05513*, 2021.

[97] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[98] Johannes Haushofer, Paul Niehaus, Carlos Paramo, Edward Miguel, and Michael W Walker. Targeting impact versus deprivation. Technical report, National Bureau of Economic Research, 2022.

[99] Jerry A Hausman, Whitney K Newey, Hidehiko Ichimura, and James L Powell. Identification and estimation of polynomial errors-in-variables models. *Journal of Econometrics*, 50(3):273–295, 1991.

[100] Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.

[101] Jennifer Hill and Yu-Sung Su. Assessing lack of common support in causal inference using bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *The Annals of Applied Statistics*, pages 1386–1420, 2013.

[102] Kevin Hillstrom. The minethatdata e-mail analytics and data mining challenge. Blog entry, Kevin Hillstrom: MineThatData, 2008.

[103] Günter J Hitsch, Sanjog Misra, and Zhang Walter. Heterogeneous treatment effects and optimal targeting policy evaluation. *Available at SSRN 3111957*, 2023.

[104] Günter J Hitsch, Sanjog Misra, and Walter Zhang. Heterogeneous treatment effects and optimal targeting policy evaluation. *Available at SSRN 3111957*, 2023.

[105] Stefan Hoderlein and Enno Mammen. Identification of marginal effects in nonseparable models without monotonicity. *Econometrica*, 75(5):1513–1518, 2007.

[106] Stefan Hoderlein and Enno Mammen. Identification and estimation of local average derivatives in non-separable models without monotonicity. *The Econometrics Journal*, 12(1):1–25, 2009.

[107] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991.

[108] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

[109] Yingyao Hu and Geert Ridder. Estimation of nonlinear models with mismeasured regressors using marginal information. *Journal of Applied Econometrics*, 27(3):347–385, 2012.

[110] Yingyao Hu and Susanne M Schennach. Instrumental variable treatment of nonclassical measurement error models. *Econometrica*, 76(1):195–216, 2008.

[111] Ta-Wei Huang and Eva Ascarza. Debiasing treatment effect estimation for privacy-protected data: A model audition and calibration approach. *Available at SSRN 4575240*, 2023.

[112] Ta-Wei Huang and Eva Ascarza. Doing more with less: Overcoming ineffective long-term targeting using short-term signals. *Marketing Science*, 0(0), 2024.

[113] Kosuke Imai and Marc Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

[114] Kosuke Imai and Aaron Strauss. Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. *Political Analysis*, 19(1):1–19, 2011.

[115] Guido Imbens, Nathan Kallus, Xiaojie Mao, and Yuhao Wang. Long-term causal inference under persistent confounding via data combination. *arXiv preprint arXiv:2202.07234*, 2022.

[116] Guido W Imbens and Whitney K Newey. Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512, 2009.

[117] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[118] Yunhun Jang, Hankook Lee, Sung Ju Hwang, and Jinwoo Shin. Learning what and where to transfer. In *International Conference on Machine Learning*, pages 3030–3039. PMLR, 2019.

[119] Andrew Jesson, Panagiotis Tigas, Joost van Amersfoort, Andreas Kirsch, Uri Shalit, and Yarin Gal. Causal-bald: Deep bayesian active learning of outcomes to infer treatment-effects from observational data. *Advances in Neural Information Processing Systems*, 34:30465–30478, 2021.

[120] Ying Jin and Shan Ba. Towards optimal variance reduction in online controlled experiments. *arXiv preprint arXiv:2110.13406*, 2021.

[121] J Morgan Jones and Jane T Landwehr. Removing heterogeneity bias from logit model estimation. *Marketing Science*, 7(1):41–59, 1988.

[122] Kousha Kalantari, Lalitha Sankar, and Anand D Sarwate. Robust privacy-utility tradeoffs under differential privacy and hamming distortion. *IEEE Transactions on Information Forensics and Security*, 13(11):2816–2830, 2018.

[123] Kathleen Kane, Victor SY Lo, and Jane Zheng. Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods. *Journal of Marketing Analytics*, 2:218–238, 2014.

[124] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

[125] Michael P Keane. Modeling heterogeneity and state dependence in consumer choice behavior. *Journal of Business & Economic Statistics*, 15(3):310–327, 1997.

[126] Edward H Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

[127] Christoph Kern, Michael P Kim, and Angela Zhou. Multi-accurate cate is robust to unknown covariate shifts. *Transactions on Machine Learning Research*, 2024.

[128] Samir Khan, Martin Saveski, and Johan Ugander. Off-policy evaluation beyond overlap: partial identification through smoothness. *arXiv preprint arXiv:2305.11812*, 2023.

[129] Patrick Kidger and Terry Lyons. Universal Approximation with Deep Narrow Networks. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2306–2327. PMLR, 09–12 Jul 2020.

[130] Michael P Kim, Christoph Kern, Shafi Goldwasser, Frauke Kreuter, and Omer Reingold. Universal adaptability: Target-independent inference that competes with propensity scoring. *Proceedings of the National Academy of Sciences*, 119(4):e2108097119, 2022.

[131] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[132] Venkataramana Kini and Ashwin Manjunatha. Revenue maximization using multitask learning for promotion recommendation. In *2020 International Conference on Data Mining Workshops (ICDMW)*, pages 144–150. IEEE, 2020.

[133] Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

[134] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

[135] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[136] Lung-fei Lee and Jungsywan H Sepanski. Estimation of linear and nonlinear errors-in-variables models using validation data. *Journal of the American Statistical Association*, 90(429):130–140, 1995.

[137] Aurélie Lemmens and Sunil Gupta. Managing churn to maximize profits. *Marketing Science*, 39(5):956–973, 2020.

[138] Tong Li and Quang Vuong. Nonparametric estimation of the measurement error model using multiple indicators. *Journal of Multivariate Analysis*, 65(2):139–165, 1998.

[139] Shiyu Liang and Rayadurgam Srikant. Why deep neural networks for function approximation? *arXiv preprint arXiv:1610.04161*, 2016.

[140] Gui Liberali and Alina Ferecatu. Morphing for consumer dynamics: Bandits meet hidden markov models. *Marketing Science*, 41(4):769–794, 2022.

[141] Sarah Lichtenstein, Baruch Fischhoff, and Lawrence D Phillips. Calibration of probabilities: The state of the art. In *Decision Making and Change in Human Affairs: Proceedings of the Fifth Research Conference on Subjective Probability, Utility, and Decision Making, Darmstadt, 1–4 September, 1975*, pages 275–324. Springer, 1977.

[142] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[143] Charles F Manski. Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4):1221–1246, 2004.

[144] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications. In *International conference on Machine learning*, pages 23803–23828. PMLR, 2023.

[145] Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

[146] Bogdan Mazoure, Paul Mineiro, Pavithra Srinath, Reza Sharifi Sedeh, Doina Precup, and Adith Swaminathan. Improving long-term metrics in recommendation systems using short-horizon reinforcement learning. *arXiv preprint arXiv:2106.00589*, 2021.

[147] Eric Mbakop and Max Tabord-Meehan. Model selection for treatment choice: Penalized welfare maximization. *Econometrica*, 89(2):825–848, 2021.

[148] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[149] Luke W Miratrix, Stefan Wager, and Jose R Zubizarreta. Shape-constrained partial identification of a population mean under unknown probabilities of sample selection. *Biometrika*, 105(1):103–114, 2018.

[150] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE, 2008.

[151] Joseph P Near, David Darais, Naomi Lefkovitz, Gary Howarth, et al. Guidelines for evaluating differential privacy guarantees. *National Institute of Standards and Technology, Tech. Rep*, 2023.

[152] Scott A Neslin, Sunil Gupta, Wagner Kamakura, Junxiang Lu, and Charlotte H Mason. Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of marketing research*, 43(2):204–211, 2006.

[153] Rachel C Nethery, Fabrizia Mealli, and Francesca Dominici. Estimating population average causal effects in the presence of non-overlap: The effect of natural gas compressor station exposure on cancer mortality. *The annals of applied statistics*, 13(2):1242, 2019.

[154] Whitney K Newey and James R Robins. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*, 2018.

[155] Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

[156] Fengshi Niu, Harsha Nori, Brian Quistorff, Rich Caruana, Donald Ngwe, and Aadharsh Kannan. Differentially private estimation of heterogeneous causal effects. In *Conference on Causal Learning and Reasoning*, pages 618–633. PMLR, 2022.

[157] Miruna Oprescu, Vasilis Syrgkanis, Keith Battocchi, Maggie Hei, and Greg Lewis. EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. https://github.com/py-why/EconML, 2019.

[158] N Padilla, E Ascarza, and O Netzer. The customer journey as a source of information. *Working Paper*, 2023.

[159] Manoranjan Pal. Consistent moment estimators of regression coefficients in the presence of errors in variables. *Journal of Econometrics*, 14(3):349–364, 1980.

[160] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. Learning model-agnostic counterfactual explanations for tabular data. In *Proceedings of The Web Conference 2020*, pages 3126–3132, 2020.

[161] Maya L Petersen, Kristin E Porter, Susan Gruber, Yue Wang, and Mark J Van Der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 21(1):31–54, 2012.

[162] Allan Pinkus. Approximation theory of the mlp model in neural networks. *Acta numerica*, 8:143–195, 1999.

[163] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[164] Ross L Prentice. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in medicine*, 8(4):431–440, 1989.

[165] Tianchen Qian, Hyesun Yoo, Predrag Klasnja, Daniel Almirall, and Susan A Murphy. Estimating time-varying causal excursion effects in mobile health with binary outcomes. *Biometrika*, 108(3):507–527, 2021.

[166] Omid Rafieian. A matrix completion solution to the problem of ignoring the ignorability assumption. 2023.

[167] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[168] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

[169] Charles S Roehrig. Conditions for identification in nonparametric and parametric models. *Econometrica*, pages 433–447, 1988.

[170] Jannik Rößler and Detlef Schoder. Bridging the gap: A systematic benchmarking of uplift modeling and heterogeneous treatment effects methods. *Journal of Interactive Marketing*, 57(4):629–650, 2022.

[171] Rishin Roy, Pradeep K Chintagunta, and Sudeep Haldar. A framework for investigating habits,"the hand of the past," and heterogeneity in dynamic brand choice. *Marketing Science*, 15(3):280–299, 1996.

[172] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

[173] Roshni Sahoo, Lihua Lei, and Stefan Wager. Learning from a biased sample. *arXiv preprint arXiv:2209.01754*, 2022.

[174] Anand D Sarwate and Lalitha Sankar. A rate-disortion perspective on local differential privacy. In *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 903–908. IEEE, 2014.

[175] Susanne M Schennach. Estimation of nonlinear models with measurement error. *Econometrica*, 72(1):33–75, 2004.

[176] Susanne M Schennach. Instrumental variable estimation of nonlinear errors-in-variables models. *Econometrica*, 75(1):201–239, 2007.

[177] Susanne M Schennach and Yingyao Hu. Nonparametric identification and semiparametric estimation of classical measurement error models without side information. *Journal of the American Statistical Association*, 108(501):177–186, 2013.

[178] David C Schmittlein, Donald G Morrison, and Richard Colombo. Counting your customers: Who-are they and what will they do next? *Management Science*, 33(1):1–24, 1987.

[179] PB Seetharaman. Modeling multiple sources of state dependence in random utility models: A distributed lag approach. *Marketing Science*, 23(2):263–271, 2004.

[180] Vira Semenova and Victor Chernozhukov. Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2):264–289, 2021.

[181] Jungsywan H Sepanski and Raymond J Carroll. Semiparametric quasilikelihood and variance function estimation in measurement error models. *Journal of Econometrics*, 58(1-2):223–256, 1993.

[182] Zhiqiang Shen, Zechun Liu, Jie Qin, Marios Savvides, and Kwang-Ting Cheng. Partial is better than all: Revisiting fine-tuning strategy for few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9594–9602, 2021.

[183] Mehrdad Showkatbakhsh, Can Karakus, and Suhas Diggavi. Privacy-utility trade-off of linear regression under random projections and additive noise. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 186–190. IEEE, 2018.

[184] Duncan Simester, Artem Timoshenko, and Spyros I Zoumpoulis. Targeting prospective customers: Robustness of machine-learning methods to typical data challenges. *Management Science*, 66(6):2495–2522, 2020.

[185] Duncan Simester, Artem Timoshenko, and Spyros I Zoumpoulis. A sample size calculation for training and certifying targeting policies. *Available at SSRN 4228297*, 2022.

[186] Adam N Smith, Stephan Seiler, and Ishant Aggarwal. Optimal price targeting. *Available at SSRN 3975957*, 2021.

[187] Liangjun Su, Takuya Ura, and Yichong Zhang. Non-separable models with high-dimensional data. *Journal of Econometrics*, 212(2):646–677, 2019.

[188] Erik Sverdrup, Ayush Kanodia, Zhengyuan Zhou, Susan Athey, and Stefan Wager. policytree: Policy learning via doubly robust empirical welfare maximization over trees. *Journal of Open Source Software*, 5(50):2232, 2020.

[189] Erik Sverdrup, Maria Petukhova, and Stefan Wager. Estimating treatment effect heterogeneity in psychiatry: A review and tutorial with causal forests. *arXiv preprint arXiv:2409.01578*, 2024.

[190] Latanya Sweeney. Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics*, 25(2-3):98–110, 1997.

[191] Artem Timoshenko, Marat Ibragimov, Duncan Simester, Jonathan Parker, and Antoinette Schoar. Transferring information between marketing campaigns to improve targeting policies. Technical report, Working Paper, 2020.

[192] Christian Toth, Lars Lorch, Christian Knoll, Andreas Krause, Franz Pernkopf, Robert Peharz, and Julius Von Kügelgen. Active bayesian causal inference. *Advances in Neural Information Processing Systems*, 35:16261–16275, 2022.

[193] Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *Advances in Neural Information Processing Systems*, 33:7852–7862, 2020.

[194] Erdem Tulin, Imai Susumu, and P Keane Michael. A model of consumer brand and quantity choice dynamics under price uncertainty. *Quantitative Marketing and Economics*, 1(1), 2002.

[195] Matilde Tullii, Solenne Gaucher, Hugo Richard, Eustache Diemert, Vianney Perchet, Alain Rakotomamonjy, Clément Calauzènes, and Maxime Vono. Position paper: Open research challenges for private advertising systems under local differential privacy. 2024.

[196] Partoo Vafaeikia, Khashayar Namdar, and Farzad Khalvati. A brief review of deep multi-task learning and auxiliary task learning. *arXiv preprint arXiv:2007.01126*, 2020.

[197] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008.

[198] Vipin Vijayachandran and R Suchithra. Iot data security by combining cryptography and local differential privacy. In *International Conference on Innovations in Cybersecurity and Data Science Proceedings of ICICDS*, pages 661–671. Springer, 2024.

[199] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[200] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

[201] Yanwen Wang, Michael Lewis, Cynthia Cryder, and Jim Sprigg. Enduring effects of goal achievement and failure within customer loyalty programs: A large-scale field experiment. *Marketing Science*, 35(4):565–575, 2016.

[202] Yuyan Wang, Mohit Sharma, Can Xu, Sriraj Badam, Qian Sun, Lee Richardson, Lisa Chung, Ed H Chi, and Minmin Chen. Surrogate for long-term user experience in recommender systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4100–4109, 2022.

[203] Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11293–11302, 2019.

[204] Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

[205] Austin Watkins, Enayat Ullah, Thanh Nguyen-Tang, and Raman Arora. Optimistic rates for multi-task representation learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[206] Justin Whitehouse, Christopher Jung, Vasilis Syrgkanis, Bryan Wilder, and Zhiwei Steven Wu. Orthogonal causal calibration. *arXiv preprint arXiv:2406.01933*, 2024.

[207] Kirk M Wolter and Wayne A Fuller. Estimation of nonlinear errors-in-variables models. *The Annals of Statistics*, pages 539–548, 1982.

[208] Xingxing Xiong, Shubo Liu, Dan Li, Zhaohui Cai, and Xiaoguang Niu. A comprehensive survey on local differential privacy. *Security and Communication Networks*, 2020(1):8829523, 2020.

[209] Ziping Xu, Amirhossein Meisami, and Ambuj Tewari. Decision making problems with funnel structure: a multi-task learning approach with application to email marketing campaigns. In *International Conference on Artificial Intelligence and Statistics*, pages 127–135. PMLR, 2021.

[210] Steve Yadlowsky, Scott Fleming, Nigam Shah, Emma Brunskill, and Stefan Wager. Evaluating treatment prioritization rules via rank-weighted average treatment effects. *arXiv preprint arXiv:2111.07966*, 2021.

[211] Steve Yadlowsky, Scott Fleming, Nigam Shah, Emma Brunskill, and Stefan Wager. Evaluating treatment prioritization rules via rank-weighted average treatment effects. *Journal of the American Statistical Association*, pages 1–14, 2024.

[212] Jeremy Yang, Dean Eckles, Paramveer Dhillon, and Sinan Aral. Targeting for long-term outcomes. *Management Science*, 2023.

[213] Mochen Yang, Edward McFowland III, Gordon Burtch, and Gediminas Adomavicius. Achieving reliable causal inference with data-mined variables: A random forest approach to the measurement error problem. *INFORMS Journal on Data Science*, 2022.

[214] Zikun Ye, Zhiqi Zhang, Dennis Zhang, Heng Zhang, and Renyu Philip Zhang. Deep learning based causal inference for large-scale combinatorial experiments: Theory and empirical evidence. *Available at SSRN 4375327*, 2023.

[215] Wei Ying, Yu Zhang, Junzhou Huang, and Qiang Yang. Transfer learning via learning to transfer. In *International Conference on Machine Learning*, pages 5085–5094. PMLR, 2018.

[216] Hema Yoganarasimhan. Search personalization using machine learning. *Management Science*, 66(3):1045–1070, 2020.

[217] Hema Yoganarasimhan, Ebrahim Barzegary, and Abhishek Pani. Design and evaluation of optimal free trials. *Management Science*, 2022.

[218] Hema Yoganarasimhan, Ebrahim Barzegary, and Abhishek Pani. Design and evaluation of optimal free trials. *Management Science*, 69(6):3220–3240, 2023.

[219] Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *ICML*, volume 1, pages 609–616, 2001.

[220] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[221] Jie Zhang and Lakshman Krishnamurthi. Customizing promotions in online stores. *Marketing science*, 23(4):561–578, 2004.

[222] Jie Zhang and Michel Wedel. The effectiveness of customized promotions in online and offline stores. *Journal of marketing research*, 46(2):190–206, 2009.

[223] Walter W Zhang and Sanjog Misra. Coarse personalization. *arXiv preprint arXiv:2204.05793*, 2022.

[224] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2021.

[225] Hao Zhong and Kaifeng Bu. Privacy-utility trade-off. *arXiv preprint arXiv:2204.12057*, 2022.

[226] Angela Yaqian Zhu, Nandita Mitra, and Jason Roy. Addressing positivity violations in causal effect estimation using gaussian process priors. *Statistics in Medicine*, 42(1):33–51, 2023.

[227] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

[228] Paul N Zivich, Jessie K Edwards, Eric T Lofgren, Stephen R Cole, Bonnie E Shook-Sa, and Justin Lessler. Transportability without positivity: a synthesis of statistical and simulation modeling. *Epidemiology*, 35(1):23–31, 2024.

Appendix A

Appendix to Chapter 1

A.1 Proofs

In this appendix we present the proofs of the theoretical results presented in the main document.

A.1.1 Positive Serial Correlation of Unexplained Variations Due to Attrition

We first prove the positive serial correlation property in Section 1.3.1. To recap, the unexplained variation for period t can be written as $\varepsilon_{i,t}^S = \mathbb{1}\left[i \text{ remains active at } t\right] \eta_{i,t}^S - \mathbb{1}\left[i \text{ has churned at } t\right] \cdot \mathbb{E}[S_{i,t}(W_i)|\mathbf{X}_i]$. Note that for $t_1 < t_2$, we have

$$\begin{aligned}
\mathbb{E}\left[\varepsilon_{i,t_1}^S \varepsilon_{i,t_2}^S\right] = & \underbrace{\theta_{t_2}(W_i|\mathbf{X}_i)\mathbb{E}\left[\eta_{i,t_1}^S\right]\mathbb{E}\left[\eta_{i,t_2}^S\right]}_{\text{If alive in both } t_1 \text{ and } t_2} + \\
& \underbrace{[\theta_{t_1}(W_i|\mathbf{X}_i) - \theta_{t_2}(W_i|\mathbf{X}_i)]\mathbb{E}\left[\eta_{i,t_1}^S\right]\mathbb{E}\left[S_{i,t_2}(W_i)|\mathbf{X}_i\right]}_{\text{If alive in } t_1 \text{ but churned in } t_2} + \\
& \underbrace{[1-\theta_{t_1}(W_i|\mathbf{X}_i)]\,\mathbb{E}\left[S_{i,t_1}(W_i)|\mathbf{X}_i\right]\mathbb{E}\left[S_{i,t_2}(W_i)|\mathbf{X}_i\right]}_{\text{If churned in } t_1}.
\end{aligned}$$

Besides, we have

$$\begin{aligned}
\mathbb{E}\left[\varepsilon_{i,t_1}^S\right]\mathbb{E}\left[\varepsilon_{i,t_2}^S\right] = & \left\{\theta_{t_1}(W_i|\mathbf{X}_i)\mathbb{E}\left[\eta_{i,t_1}^S\right] - [1-\theta_{t_1}(W_i|\mathbf{X}_i)]\,\mathbb{E}\left[S_{i,t_1}(W_i)|\mathbf{X}_i\right]\right\} \times \\
& \left\{\theta_{t_2}(W_i|\mathbf{X}_i)\mathbb{E}\left[\eta_{i,t_2}^S\right] - [1-\theta_{t_2}(W_i|\mathbf{X}_i)]\,\mathbb{E}\left[S_{i,t_2}(W_i)|\mathbf{X}_i\right]\right\}.
\end{aligned}$$

Combining this two terms, we can derive the covariance:

$$\begin{aligned}\text{Cov}\left[\varepsilon^{\theta}_{i,t_1}, \varepsilon^{\theta}_{i,t_2}\right] &= \mathbb{E}\left[\varepsilon^{S}_{i,t_1}\varepsilon^{S}_{i,t_2}\right] - \mathbb{E}\left[\varepsilon^{S}_{i,t_1}\right]\mathbb{E}\left[\varepsilon^{S}_{i,t_2}\right] \\ &= \left[1-\theta_{t_1}(W_i|\mathbf{X}_i)\right]\left\{\mathbb{E}\left[S_{i,t_1}(W_i)|\mathbf{X}_i\right] + \mathbb{E}\left[\eta^{S}_{i,t_1}\right]\right\}\theta_{t_2}(W_i|\mathbf{X}_i)\left\{\mathbb{E}\left[S_{i,t_2}(W_i)|\mathbf{X}_i\right] + \mathbb{E}\left[\eta^{S}_{i,t_2}\right]\right\} \geqslant 0.\end{aligned}$$

since every term in the above equation is larger than zero.

A.1.2 Proof for Theorem 1

Class of CATE Estimation.

We start by describe the class of CATE estimators we consider here in the theoretical analysis.

Assumption A.1 ***(Class of CATE Estimator)*** *For a given individual with covariates* $\mathbf{x}_{\text{new}}$*, the predicted CATE* $\widehat{\tau}(\mathbf{x}_{\text{new}}|\mathcal{D}^o, \mathcal{D}^\ell, \mathcal{D}^m)$ *can be expressed as the difference between two (adjusted) outcome estimators,* $\widehat{\mu}^0$ *and* $\widehat{\mu}^1$*, in the following form:*

$$\widehat{\mu}^w(\mathbf{x}_{\text{new}}) = \sum_{i\in\mathcal{I}^o: W_i = w} \widehat{\ell}^w_i(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)[Y_i - \widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)],$$

where $\mathcal{I}^o$ *denotes the set of individuals used to impute the two outcome predictions,* $\mathcal{D}^\ell = \{\mathcal{X}^\ell, \mathcal{Y}^\ell, \mathcal{W}^\ell\}$ *represents the data used to construct the weight function* $\widehat{\ell}^w_i(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)$*, and* $\mathcal{D}^m = \{\mathcal{X}^m, \mathcal{Y}^m, \mathcal{W}^m\}$ *is the data used to determine the adjustment function* $\widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)$*. We also denote* $\mathcal{D}^o = \{\mathcal{X}^o, \mathcal{Y}^o, \mathcal{W}^o\}$ *as the experiment data of individuals in* $\mathcal{I}_o$*, Furthermore, we assume that the estimation process of the CATE model satisfy the following conditions:*

1. *(Honest Estimation) The weight function for individual* j *is independent of* Y_j *given* $\mathbf{X}_i$*,* $\forall j \in \mathcal{I}^o$*.*
2. *(Cross Fitting) The adjustment function for individual* j *is either zero or constructed from the dataset* $\mathcal{D}^m$ *that is independent of* $(\mathbf{X}_j, Y_j)$*.*

For the theoretical analysis, we also assume that adjustment function, $\widehat{m}^w(\mathbf{x}|\mathcal{D}^m)$*, can be written as* $\widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m) = \sum_{j\in\mathcal{D}^m} S_j(\mathbf{x})Y_j$*, where* S_j *is (i) independent of any outcome information or (ii) can be represented as the weight of a potential n-nearest-neighbors estimator as described in Definition 7 in [199]. Note that (i) includes methods such as kernel regression, while (ii) is specifically relevant for tree-based methods like Causal Forest.*

Next, we show that the class of CATE estimators include a wide variety of commonly-used CATE models, such as S-learner, T-learner, Causal Forest, and R-learner.

Proposition A.1 *The following CATE estimators belong to the class described in Assumption A.1:*

1. *S-learners and T-learners [134] with the outcome models being OLS regressions, ridge regressions, k-nearest neighbors, or random forests with honest estimation.*

2. *R-learners [155] with the CATE predictor and the first-stage conditional mean model being OLS regressions, ridge regressions, k-nearest neighbors, or random forests with honest estimation .*

3. *Causal Forest with honest estimation.*

Proof: Let $\mathcal{D} = \{(\mathbf{X}_i, W_i, Y_i)\}_{i=1}^N$ be the training set for CATE prediction.

1. S-learners and T-learners:

We provide a proof for S-learners using (a) high-dimensional ridge regression (the proof for OLS estimators is similar), (b) k-nearest neighbors, and (c) random forests with honest estimation as the outcome model. We omit the proofs for T-learner as they are similar to the proofs for S-learner.

(a) High-dimensional Ridge Regression:

Consider the ridge regression model $Y_i = \phi(\mathbf{X}_i, W_i)'\boldsymbol{\beta} + \varepsilon_i$, where $\phi(\mathbf{X}_i, W_i)$ is a high-dimensional feature transformation function used to construct a ridge regression estimator.

Let $\Phi = [\phi(\mathbf{X}_1, W_1) \cdots \phi(\mathbf{X}_N, W_N)]'$ denote the feature matrix. Then, the closed-form solution of the ridge coefficient with regularization term λ can be written as:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{\Phi}'\boldsymbol{\Phi} + \lambda I)^{-1}\boldsymbol{\Phi}'\mathbf{y}.$$

Denote $\mathbf{p}_i$ be the i-th row of the matrix $(\boldsymbol{\Phi}'\boldsymbol{\Phi} + \lambda I)^{-1}\boldsymbol{\Phi}'$. Then, the predicted CATE for $\mathbf{x}_{\text{new}}$ can be written as

$$\begin{aligned}
\widehat{\tau}_Y(\mathbf{x}_{\text{new}}) &= \sum_{i\in\mathcal{D}} \left[\phi(\mathbf{x}_{\text{new}}, 1) - \phi(\mathbf{x}_{\text{new}}, 0)\right]' \mathbf{p}_i Y_i \\
&= \sum_{i\in\mathcal{D}:\ W_i=1} \underbrace{\left[\phi(\mathbf{x}_{\text{new}}, 1) - \phi(\mathbf{x}_{\text{new}}, 0)\right]' \mathbf{p}_i}_{\widehat{\ell}_i^1(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)} Y_i - \sum_{i\in\mathcal{D}:\ W_i=0} \underbrace{\left[\phi(\mathbf{x}_{\text{new}}, 0) - \phi(\mathbf{x}_{\text{new}}, 1)\right]' \mathbf{p}_i}_{\widehat{\ell}_i^0(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)} Y_i.
\end{aligned}$$

Note that this formulation satisfies Assumption A.1 with $\mathcal{D}^o = \mathcal{D}^\ell = \mathcal{D}$ and $\widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m) \equiv 0$. Specefically, the honest estimation assumption is satisfied since $\mathbf{p}_i$ only depends on the covariates of customer i.

(b) k-nearest neighbors:

Let $N_k(\mathbf{x}_{\text{new}}, w)$ represent the set of the k-nearest neighboring customers in the training set for the new customer with covariates and treatment assignment $(\mathbf{x}_{\text{new}}, w)$. Then, we can express the predicted CATE as:

$$\begin{aligned}
\widehat{\tau}_Y(\mathbf{x}_{\text{new}}) &= \sum_{\mathbf{X}_i \in N_k(\mathbf{x}_{\text{new}}, 1)} \frac{Y_i}{k} - \sum_{\mathbf{X}_i \in N_k(\mathbf{x}_{\text{new}}, 0)} \frac{Y_i}{k} \\
&= \sum_{i \in \mathcal{D}:\ W_i = 1} \underbrace{\frac{\mathbb{1}[\mathbf{X}_i \in N_k(\mathbf{x}_{\text{new}}, 1)] - \mathbb{1}[\mathbf{X}_i \in N_k(\mathbf{x}_{\text{new}}, 0)]}{k}}_{\widehat{\ell}_i^1(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)} Y_i - \\
&\qquad \sum_{i \in \mathcal{D}:\ W_i = 0} \underbrace{\frac{\mathbb{1}[\mathbf{X}_i \in N_k(\mathbf{x}_{\text{new}}, 0)] - \mathbb{1}[\mathbf{X}_i \in N_k(\mathbf{x}_{\text{new}}, 1)]}{k}}_{\widehat{\ell}_i^0(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)} Y_i.
\end{aligned}$$

Since the weight only depends on the covariates of customer i, it satisfies Assumption A.1 with $\mathcal{D}^o = \mathcal{D}^\ell = \mathcal{D}$ and $\widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m) \equiv 0$.

(c) Random Forest with Honest Estimation:

Let $\mathcal{D}_1^b, \mathcal{D}_2^b$ be the divided samples for the b-th tree ($b = 1, \cdots, B$), where $\mathcal{D}_1^b$ is used to construct the regression tree and $\mathcal{D}_2^b$ is used to generate predictions. Define $L^b(\mathbf{x}_{\text{new}}, W)$ be the leaf in the b-th tree to which customer $(\mathbf{x}_{\text{new}}, W)$ belongs. Using this notation, we can express the predicted CATE as follows:

$$\begin{aligned}
\widehat{\tau}_Y(\mathbf{x}_{\text{new}}) &= \sum_{i=1}^{N} \frac{1}{B} \sum_{b=1}^{B} \left[\frac{\mathbb{1}[\mathbf{X}_i \in L^b(\mathbf{x}_{\text{new}}, 1),\ i \in \mathcal{D}_2^b]}{|\{i \in \mathcal{D}_2^b :\ \mathbf{X}_i \in L^b(\mathbf{x}_{\text{new}}, 1)\}|} - \frac{\mathbb{1}[\mathbf{X}_i \in L^b(\mathbf{x}_{\text{new}}, 0),\ i \in \mathcal{D}_2^b]}{|\{i \in \mathcal{D}_2^b :\ \mathbf{X}_i \in L^b(\mathbf{x}_{\text{new}}, 0)\}|} \right] Y_i \\
&= \sum_{i \in \mathcal{D}: W_i = 1} \underbrace{\frac{1}{B} \sum_{b=1}^{B} \left(\frac{\mathbb{1}[\mathbf{X}_i \in L^b(\mathbf{x}_{\text{new}}, 1),\ i \in \mathcal{D}_2^b]}{|\{i \in \mathcal{D}_2^b :\ \mathbf{X}_i \in L^b(\mathbf{x}_{\text{new}}, 1)\}|} - \frac{\mathbb{1}[\mathbf{X}_i \in L^b(\mathbf{x}_{\text{new}}, 0),\ i \in \mathcal{D}_2^b]}{|\{i \in \mathcal{D}_2^b :\ \mathbf{X}_i \in L^b(\mathbf{x}_{\text{new}}, 0)\}|} \right)}_{\widehat{\ell}_i^1(\mathbf{x}_{\text{new}})} Y_i - \\
&\qquad \sum_{i \in \mathcal{D}: W_i = 0} \underbrace{\frac{1}{B} \sum_{b=1}^{B} \left(\frac{\mathbb{1}[\mathbf{X}_i \in L^b(\mathbf{x}_{\text{new}}, 0),\ i \in \mathcal{D}_2^b]}{|\{i \in \mathcal{D}_2^b :\ \mathbf{X}_i \in L^b(\mathbf{x}_{\text{new}}, 0)\}|} - \frac{\mathbb{1}[\mathbf{X}_i \in L^b(\mathbf{x}_{\text{new}}, 1),\ i \in \mathcal{D}_2^b]}{|\{i \in \mathcal{D}_2^b :\ \mathbf{X}_i \in L^b(\mathbf{x}_{\text{new}}, 0)\}|} \right)}_{\widehat{\ell}_i^0(\mathbf{x}_{\text{new}})} Y_i.
\end{aligned}$$

Note that the weight function is independent of Y_i due to the honest estimation. Therefore, the

above formulation satisfies Assumption A.1

2. R-learners:

We only present the scenario where high-dimensional ridge regression is used in the second-stage estimation, as the derivations for k-nearest neighbors and random forests with honest estimation are similar, using the weights derived in 1.(b) and 1.(c).

Let $\mathcal{D}^1_{\text{train}}, \cdots, \mathcal{D}^Q_{\text{train}}$ be the sample splits of the training set. Define $\mathcal{D}^{(-i)}_{\text{train}} = \cup_{q:\ i \notin \mathcal{D}^q_{\text{train}}} \mathcal{D}^q_{\text{train}}$ as the training set that excludes the subsample that includes the i-th sample. Also, denote $\widehat{e}^{\mathcal{D}^{(-i)}_{\text{train}}}(\mathbf{X}_i)$ (for propensity scores) and $\widehat{m}^{\mathcal{D}^{(-i)}}(\mathbf{X}_i)$ (for conditional means) as the nuisance models trained using the data set $\mathcal{D}^{(-i)}_{\text{train}}$. Then, the Robinson's transformation for each customer is

$$\widetilde{\tau}_i = \frac{Y_i - \widehat{m}^{\mathcal{D}^{(-i)}_{\text{train}}}(\mathbf{X}_i)}{W_i - \widehat{e}^{\mathcal{D}^{(-i)}_{\text{train}}}(\mathbf{X}_i)}.$$

Now, consider the ridge regression model $\widetilde{\tau}_i = \phi(\mathbf{x}_i)'\boldsymbol{\beta} + \varepsilon_i$, where $\phi(\mathbf{x})$ is the high-dimensional feature transformation function. Let $\Phi = [\phi(\mathbf{x}_1) \ \cdots \ \phi(\mathbf{x}_N)]'$ be the feature matrix. Then, the closed-form solution for the ridge coefficient is

$$\widehat{\beta} = (\boldsymbol{\Phi}'\boldsymbol{\Phi} + \lambda I)^{-1}\boldsymbol{\Phi}'\widetilde{\boldsymbol{\tau}}.$$

Denote $\mathbf{p}_i$ be the i-th row of the matrix $(\boldsymbol{\Phi}'\boldsymbol{\Phi} + \lambda I)^{-1}\boldsymbol{\Phi}'$. Then, the predicted CATE for $\mathbf{x}_{\text{new}}$ can be written as

$$\begin{aligned}
\widehat{\tau}_Y(\mathbf{x}_{\text{new}}) &= \sum_{i \in \mathcal{D}} \phi(\mathbf{x}_{\text{new}})'\mathbf{p}_i \left(\frac{Y_i - \widehat{m}^{\mathcal{D}^{(-i)}_{\text{train}}}(\mathbf{X}_i)}{W_i - \widehat{e}^{\mathcal{D}^{(-i)}_{\text{train}}}(\mathbf{X}_i)} \right) \\
&= \sum_{i \in \mathcal{D}: W_i = 1} \underbrace{\left(\frac{\phi(\mathbf{x}_{\text{new}})'\mathbf{p}_i}{1 - \widehat{e}^{\mathcal{D}^{(-i)}_{\text{train}}}(\mathbf{X}_i)} \right)}_{\widehat{\ell}^1_i(\mathbf{x}_{\text{new}})} [Y_i - \underbrace{\widehat{m}^{\mathcal{D}^{(-i)}_{\text{train}}}(\mathbf{X}_i)}_{\widehat{m}^1(\mathbf{X}_i)}] - \sum_{i \in \mathcal{D}: W_i = 0} \underbrace{\left(\frac{\phi(\mathbf{x}_{\text{new}})'\mathbf{p}_i}{1 - \widehat{e}^{\mathcal{D}^{(-i)}_{\text{train}}}(\mathbf{X}_i)} \right)}_{\widehat{\ell}^0_i(\mathbf{x}_{\text{new}})} [Y_i - \underbrace{\widehat{m}^{\mathcal{D}^{(-i)}_{\text{train}}}(\mathbf{X}_i)}_{\widehat{m}^0(\mathbf{X}_i)}].
\end{aligned}$$

Note that $\widehat{\ell}^{W_i}_i(\mathbf{x}_{\text{new}})$ is independent of Y_i as it only depends on the covariate information. Also, $\widehat{m}^w(\mathbf{X}_i)$ is independent of Y_i and $\widehat{\ell}^{W_i}_i(\mathbf{x}_{\text{new}})$ as it is estimated using $\mathcal{D}^{(-i)}_{\text{train}}$. Therefore, it satisfies Assumption A.1.

3. Causal Forest:

Let $\mathcal{D}^b_1, \mathcal{D}^b_2$ be the divided samples for the b-th tree ($b = 1, \cdots, B$), where $\mathcal{D}^b_1$ is used to construct the causal tree and $\mathcal{D}^b_2$ is used to generate predictions. Define $L^b(\mathbf{x}_{\text{new}}$ be the leaf in the b-th tree

to which customer $(\mathbf{x}_{\text{new}}, W)$ belongs. Using this notation, we can express the predicted CATE as follows:

$$\widehat{\tau}_Y(\mathbf{x}_{\text{new}}) = \sum_{i\in\mathcal{D}:W_i=1} \underbrace{\frac{1}{B}\sum_{b=1}^{B} \frac{\mathbb{1}[\mathbf{X}_i \in L^b(\mathbf{x}_{\text{new}}),\ i\in\mathcal{D}_2^b]}{|\{i\in\mathcal{D}_2^b:\ \mathbf{X}_i\in L^b(\mathbf{x}_{\text{new}})\}|}}_{\widehat{\ell}_i^1(\mathbf{x}_{\text{new}})} Y_i - \sum_{i\in\mathcal{D}:W_i=0} \underbrace{\frac{1}{B}\sum_{b=1}^{B} \frac{\mathbb{1}[\mathbf{X}_i \in L^b(\mathbf{x}_{\text{new}}),\ i\in\mathcal{D}_2^b]}{|\{i\in\mathcal{D}_2^b:\ \mathbf{X}_i\in L^b(\mathbf{x}_{\text{new}})\}|}}_{\widehat{\ell}_i^0(\mathbf{x}_{\text{new}})} Y_i.$$

Note that $\widehat{\ell}_i^{W_i}(\mathbf{x}_{\text{new}})$ is independent of Y_i due to the honest estimation procedure in Causal Forest. Therefore, it satisfies Assumption A.1.

Proof for Theorem 1.1.

Let $\varepsilon_i^{Y_T} = Y_{i,T} - \mathbb{E}[Y_{i,T}(W_i)|\mathbf{X}_i]$ be the variations in $Y_{i,T}$ that cannot be explained by $\mathbf{X}_i$ and W_i, and denote $\sigma^2 = \text{Var}[\varepsilon_i^{Y_T}]$.

Since the experiment data are i.i.d. samples, the variance of the predicted CATEs can be written as

$$\begin{aligned}
\text{Var}\left[\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}}) \,\middle|\, \{\mathbf{X}_i, W_i\}_{i=1}^N\right] &= \sum_{i\in\mathcal{I}^o:W_i=1} \text{Var}\left\{\widehat{\ell}_i^1(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)[Y_{i,T} - \widehat{m}^1(\mathbf{X}_i|\mathcal{D}^m) \,\middle|\, \{\mathbf{X}_i, W_i\}_{i=1}^N\right\} + \\
&\quad \sum_{i\in\mathcal{I}^o:W_i=0} \text{Var}\left\{\widehat{\ell}_i^0(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)[Y_{i,T} - \widehat{m}^0(\mathbf{X}_i|\mathcal{D}^m) \,\middle|\, \{\mathbf{X}_i, W_i\}_{i=1}^N\right\} \\
&= \sum_{i\in\mathcal{I}^o:W_i=1} \text{Var}\left\{\widehat{\ell}_i^1(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)[\varepsilon_i^{Y_T} - \widehat{m}^1(\mathbf{X}_i|\mathcal{D}^m)\right\} + \\
&\quad \sum_{i\in\mathcal{I}^o:W_i=0} \text{Var}\left\{\widehat{\ell}_i^0(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)[\varepsilon_i^{Y_T} - \widehat{m}^0(\mathbf{X}_i|\mathcal{D}^m)\right\}
\end{aligned}$$

since $\mathbb{E}[Y_{i,T}(W_i)|\mathbf{X}_i]$ is a constant given $(\mathbf{X}_i, W_i)$.

For each individual i, we can write its variance as

$$\begin{aligned}
&\text{Var}\{\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)[\varepsilon_i^{Y_T} - \widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m])\} \\
&= \mathbb{E}\left\{[\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)]^2\right\}\left(\text{Var}[\varepsilon_i^{Y_T} - \widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)]\right) + \text{Var}[\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)]\mathbb{E}^2[\varepsilon_i^{Y_T} - \widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)] \\
&= \mathbb{E}\left\{[\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)]^2\right\}\left(\sigma^2 + \text{Var}[\widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)]\right) + \text{Var}[\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)]\left\{\sigma^2 + \mathbb{E}^2[\widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)]\right\}.
\end{aligned}$$

Now, our goal is to bound the variance by showing:

1. the variance of $\widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m) = \Theta(\sigma^2)$, where Θ is the asymptotic notation[1],

2. the expectation of the adjustment function $\mathbb{E}^2\left[\widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)\right]$ do not scale with σ^2, and

3. $\mathbb{E}\left\{[\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)]^2\right\}$ and $\text{Var}[\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)]$ do not scale with σ^2.

Proof for (a): To show (a), note that the variance of the adjustment function can be written as:

$$\begin{aligned}
\text{Var}[\widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)] &= \sum_{j\in\mathcal{D}^m} \text{Var}[S_j(\mathbf{X}_i)Y_{j,T}] \\
&= \sum_{j\in\mathcal{D}^m} \text{Var}\left\{S_j(\mathbf{X}_i)\mathbb{E}[Y_{j,T}|\mathbf{X}_j]\right\} + \sum_{j\in\mathcal{D}^m} \text{Var}\left[S_j(\mathbf{X}_i)\varepsilon_j^{Y_T}\right] \\
&= \underbrace{\sum_{j\in\mathcal{D}^m} \text{Var}\left\{S_j(\mathbf{X}_i)\mathbb{E}[Y_{j,T}|\mathbf{X}_j]\right\}}_{\equiv C_1} + \sigma^2 \underbrace{\sum_{j\in\mathcal{D}^m} \left\{\text{Var}\left[S_j(\mathbf{X}_i)\right] + \mathbb{E}^2\left[S_j(\mathbf{X}_i)\right]\right\}}_{\equiv C_2}.
\end{aligned}$$

The last equation hold because $\text{Var}[XY] = \text{Var}[X]\text{Var}[Y] + \text{Var}[Y]\mathbb{E}^2[X] + \text{Var}[X]\mathbb{E}^2[Y]$.

Now, our goal is to show that C_1 and C_2 do not scale with σ^2. In the case when $S_j(\mathbf{X}_i)$ is independent of the outcome (and therefore $\varepsilon_j^{Y_T}$), $S_j(\mathbf{X}_i)$ does not depend on σ^2 by construction. Therefore, C_1 and C_2 do not scale with σ^2. If $\widehat{m}$ is a potential n-nearest-neighbors estimator, by the fact from the discussion of Lemma 4 in [199], we have

$$\frac{1}{2(n-1)|\mathcal{D}^m|} \lesssim \text{Var}[S_j(\mathbf{X}_i)] \lesssim \frac{1}{(n-1)|\mathcal{D}^m|} \quad \text{and} \quad 0 \leqslant \mathbb{E}[S_j(\mathbf{X}_i)] \leqslant 1.$$

In this case, C_1 and C_2 do not scale with σ^2. Therefore, we can conclude that $\text{Var}[\widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)] = \Theta(\sigma^2)$.

Proof for (b): If the weight of $\widehat{m}$ only depends on covariates, the expectation can be written as

$$\mathbb{E}[\widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)] = \sum_{j\in\mathcal{D}^m} \mathbb{E}[S_j(\mathbf{X}_i)Y_{j,T}(W_j)] = \sum_{j\in\mathcal{D}^m} \mathbb{E}[S_j(\mathbf{X}_i)]\mathbb{E}[Y_{j,T}(W_j)|\mathbf{X}_j].$$

Therefore, $\mathbb{E}[\widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)]$ do not depend on σ^2.

[1]The definition of Θ-function is as follows: a function $f(\sigma^2) = \Theta(\sigma^2)$ if there exists σ_0^2 and C_1, C_2 such that $C_1\sigma^2 \leqslant f(\sigma^2) \leqslant C_2\sigma^2$ for any $\sigma^2 > \sigma_0^2$.

If $\widehat{m}$ is a potential n-nearest-neighbors estimator, since $0 \leqslant S_j(\mathbf{X}_i) \leqslant 1$, we have

$$-\sum_{j\in\mathcal{D}^m} |\mathbb{E}[Y_{j,T}(W_j)|\mathbf{X}_j]| \leqslant \mathbb{E}[\widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)] \leqslant \sum_{j\in\mathcal{D}^m} |\mathbb{E}[Y_{j,T}(W_j)|\mathbf{X}_j]|.$$

Since the upper bound and lower bound of $\mathbb{E}[\widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)]$ do not depend on σ^2, we can conclude that $\mathbb{E}^2\left[\widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)\right]$ do not scale with σ^2.

Proof for (c): When the weight function only depends on the covariate information (e.g., CATE models in Proposition A.1 other than Causal Forest and R-learner with the honest random forest estimator in the second stage), $\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)$ do not scale with σ^2 by nature.

When the weight can be viewed as a potential n-nearest neighbors weight (e.g., Causal Forest and R-learner with the honest random forest estimator in the second stage), we have

$$\frac{1}{2(n-1)|\mathcal{D}^\ell|} \lesssim \text{Var}[\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)] \lesssim \frac{1}{(n-1)|\mathcal{D}^\ell|} \quad \text{and} \quad 0 \leqslant \mathbb{E}[\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)] \leqslant \frac{1}{|\mathcal{D}^\ell|}.$$

As a result, $\mathbb{E}\left\{[\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)]^2\right\}$ and $\text{Var}[\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)]$ do not scale with σ^2.

A.1.3 Proof for Corollary 1.1

First, consider the case when the true CATE $\tau_{Y_T}(\mathbf{x}_{\text{new}}) > 0$. Then, the probability that the learned policy makes a different decision from the optimal targeting policy is

$$\begin{aligned}
&\mathbb{P}\left[\tau_{Y_T}(\mathbf{x}_{\text{new}}) \cdot \widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}}) < 0\right] \\
&\quad = \mathbb{P}\left[\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}}) < 0\right] = \mathbb{P}\left[\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}}) - \tau_{Y_T}(\mathbf{x}_{\text{new}}) < -\tau_{Y_T}(\mathbf{x}_{\text{new}})\right] \\
&\quad = \mathbb{P}\left[\underbrace{\frac{\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}}) - \tau_{Y_T}(\mathbf{x}_{\text{new}})}{\sqrt{\text{Var}\left[\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}})\right]}}}_{\equiv Z} < \frac{-\tau_{Y_T}(\mathbf{x}_{\text{new}})}{\sqrt{\text{Var}\left[\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}})\right]}}\right] = F_Z\left(\frac{-\tau_{Y_T}(\mathbf{x}_{\text{new}})}{\sqrt{\text{Var}\left[\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}})\right]}}\right),
\end{aligned}$$

where $Z \equiv \frac{\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}}) - \tau_{Y_T}(\mathbf{x}_{\text{new}})}{\sqrt{\text{Var}\left[\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}})\right]}}$ has mean zero and variance one, and F_Z is the cumulative distribution function of Z. Since (i) $\frac{-\tau_{Y_T}(\mathbf{x}_{\text{new}})}{\sqrt{\text{Var}\left[\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}})\right]}}$ is increasing in $\sqrt{\text{Var}\left[\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}})\right]}$ and (ii) F_Z is a weakly increasing function, we can conclude that the mistarteting probability is weakly increasing in $\sqrt{\text{Var}\left[\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}})\right]}$.

Similarly, for the case when the true $\tau_{Y_T}(\mathbf{x}_{\text{new}}) < 0$, we have

$$
\begin{aligned}
&\mathbb{P}\left[\tau_{Y_T}(\mathbf{x}_{\text{new}}) \cdot \widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}}) < 0\right] \\
&\quad = \mathbb{P}\left[\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}}) > 0\right] = \mathbb{P}[\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}}) - \tau_{Y_T}(\mathbf{x}_{\text{new}}) > \underbrace{-\tau_{Y_T}(\mathbf{x}_{\text{new}})}_{=|\tau_{Y_T}(\mathbf{x}_{\text{new}})|}] \\
&\quad = \mathbb{P}\left[\underbrace{\frac{\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}}) - \tau_{Y_T}(\mathbf{x}_{\text{new}})}{\sqrt{\text{Var}\left[\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}})\right]}}}_{\equiv Z} > \frac{|\tau_{Y_T}(\mathbf{x}_{\text{new}})|}{\sqrt{\text{Var}\left[\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}})\right]}}\right] \\
&\quad = 1 - \mathbb{P}\left[Z < \frac{|\tau_{Y_T}(\mathbf{x}_{\text{new}})|}{\sqrt{\text{Var}\left[\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}})\right]}}\right] = 1 - F_Z\left(\frac{|\tau_{Y_T}(\mathbf{x}_{\text{new}})|}{\sqrt{\text{Var}\left[\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}})\right]}}\right),
\end{aligned}
$$

Note that $F_Z\left(\frac{|\tau_{Y_T}(\mathbf{x}_{\text{new}}|)}{\sqrt{\text{Var}\left[\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}})\right]}}\right)$ is weakly decreasing in $\sqrt{\text{Var}\left[\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}})\right]}$ since (i) $\frac{|\tau_{Y_T}(\mathbf{x}_{\text{new}})|}{\sqrt{\text{Var}\left[\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}})\right]}}$ becomes smaller when $\sqrt{\text{Var}\left[\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}})\right]}$ becomes larger and (ii) F_Z is a weakly increasing function. Therefore, the mistargeting probability is weakly increasing in $\sqrt{\text{Var}\left[\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}})\right]}$.

Now, let us consider the case when the firm aims to target customers with CATEs larger than the threshold c. Following the same approach, we can write the mistargeting probability is as:

$$
\begin{aligned}
&\mathbb{P}\left[(\tau_{Y_T}(\mathbf{x}_{\text{new}}) - c) \cdot (\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}}) - c) < 0\right] = \\
&\begin{cases}
\mathbb{P}\left[\frac{\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}}) - \tau_{Y_T}(\mathbf{x}_{\text{new}})}{\sqrt{\text{Var}\left[\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}})\right]}} < \frac{-|\tau_{Y_T}(\mathbf{x}_{\text{new}}) - c|}{\sqrt{\text{Var}\left[\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}})\right]}}\right], & \text{if } \tau_{Y_T}(\mathbf{x}_{\text{new}}) - c > 0, \\
1 - \mathbb{P}\left[\frac{\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}}) - \tau_{Y_T}(\mathbf{x}_{\text{new}})}{\sqrt{\text{Var}\left[\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}})\right]}} < \frac{|\tau_{Y_T}(\mathbf{x}_{\text{new}}) - c|}{\sqrt{\text{Var}\left[\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}})\right]}}\right], & \text{if } \tau_{Y_T}(\mathbf{x}_{\text{new}}) - c < 0.
\end{cases}
\end{aligned}
$$

Following the above analysis, it is clear that the mistargeting probability increases as $\text{Var}\left[\widehat{\tau}_{Y_T}(\mathbf{x}_{\text{new}})\right]$ increases.

A.2 Implications of Multiplicative Noise Structure

In this appendix, we first use a simple example to elucidate how a surrogate model may result in higher variance for surrogate models when the outcome variable is a product of two distinct variables. We then showcase the advantages of the separate imputation approach in terms of variance reduction.

A.2.1 Illustration Using Simple Linear Regressions

Let's consider the following data generating process:

$$\begin{aligned} \mathcal{T}_i &= S_i + \varepsilon_i^{\mathcal{T}}, \qquad \Lambda_i = S_i + \varepsilon_i^{\Lambda}, \\ Y_i &\equiv \mathcal{T}_i \times \Lambda_i = S_i^2 + \underbrace{(\varepsilon_i^{\Lambda} + \varepsilon_i^{\mathcal{T}})S_i}_{\text{Additional Heterogeneity}} + \varepsilon_i^{\mathcal{T}}\varepsilon_i^{\Lambda}, \end{aligned} \tag{A.1}$$

where $\varepsilon_i^{\mathcal{T}} \sim \mathcal{N}(0, \sigma_{\mathcal{T}}^2)$ and $\varepsilon_i^{\Lambda} \sim \mathcal{N}(0, \sigma_{\Lambda}^2)$ are independent. Essentially, the multiplicative structure result in a random effect $(\varepsilon_i^{\Lambda} + \varepsilon_i^{\mathcal{T}})S_i$ that varies across individual.

Now consider the "single" imputation approach where we fit a linear regression model $Y_i = \gamma_0 + \gamma_1 S_i^2 + \eta_i$ using least-square estimation. Using the standard result, we have

$$\widehat{\gamma}_1 = 1 + \frac{\sum_i (S_i^2 - \overline{S^2}) \left[\left(\varepsilon_i^{\Lambda} + \varepsilon_i^{\mathcal{T}}\right) S_i + \varepsilon_i^{\mathcal{T}}\varepsilon_i^{\Lambda} \right]}{\sum_i \left(S_i^2 - \overline{S^2}\right)^2}.$$

Note that (i) $\widehat{\gamma}_1$ is an unbiased estimator, i.e., $\mathbb{E}[\widehat{\gamma}_1] = 1$, and (ii) the variance of $\widehat{\gamma}_1$ is

$$\text{Var}(\widehat{\gamma}_1) = \frac{1}{\left[\sum_i \left(S_i^2 - \overline{S^2}\right)^2\right]^2} \sigma_{\Lambda}^2 \sigma_{\mathcal{T}}^2 + \frac{\sum_i S_i^2 \left(S_i^2 - \overline{S^2}\right)^2}{\left[\sum_i \left(S_i^2 - \overline{S^2}\right)^2\right]^2} \left[\sigma_{\Lambda}^2 + \sigma_{\mathcal{T}}^2\right].$$

The first term in the above equation is the variance of the estimated coefficient if there is no random effect $(\varepsilon_i^{\Lambda} + \varepsilon_i^{\mathcal{T}})S_i$. The second term captures the additional variance caused by the additional heterogeneity term $((\varepsilon_i^{\Lambda} + \varepsilon_i^{\mathcal{T}})S_i)$. This simple analysis shows that the multiplicative structure can increase the variance of the surrogate model.

Next, we consider the "separate imputation" approach, where we fit two regression models $\mathcal{T}_i = \beta_0 + \beta_1 S_i + \varepsilon_i^{\mathcal{T}}$ and $\Lambda_i = \delta_0 + \delta_1 S_i + \varepsilon_i^{\Lambda}$. In this case, we have

$$\text{Var}(\widehat{\beta}_1) = \frac{1}{\sum_i \left(S_i - \overline{S}\right)^2} \sigma_{\mathcal{T}}^2, \qquad \text{Var}(\widehat{\delta}_1) = \frac{1}{\sum_i \left(S_i - \overline{S}\right)^2} \sigma_{\Lambda}^2.$$

In this case, both variance terms are not affected by $(\varepsilon_i^{\Lambda} + \varepsilon_i^{\mathcal{T}})S_i$. Therefore, the separate imputation approach has much smaller variance compared to the single imputation model.

A.2.2 Simulation Evidence

Next, we turn to simulation evidence to underscore the variance reduction benefit of the proposed separate imputation approach. We generate 100 sets of historical data ($d = 1, \cdots, 100$), each comprising 500 customers, in accordance with the data generation process detailed in (A.1), where $S_i \sim \mathcal{N}(0, 1)$. For each set, we implement four different approaches:

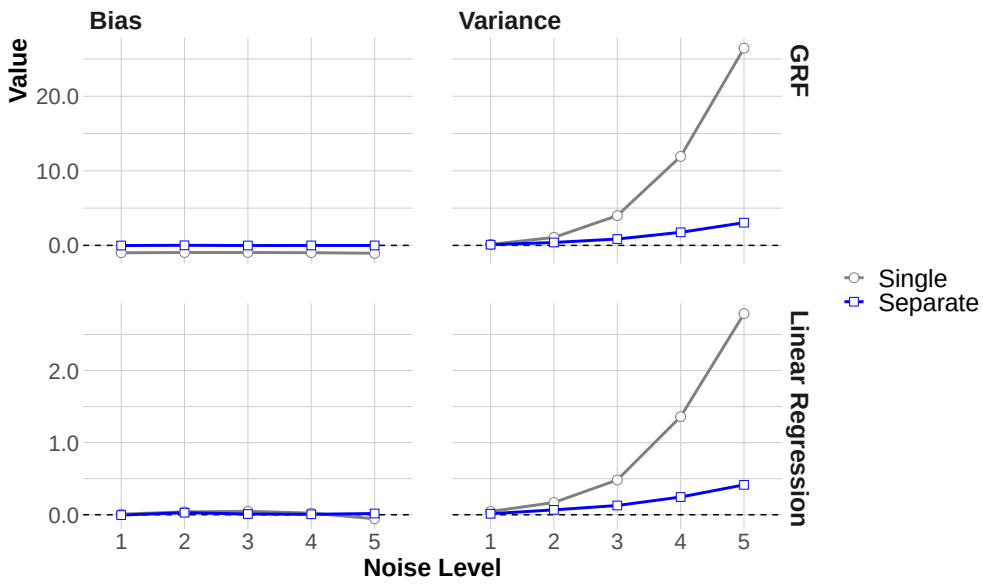
1. Single linear regression: $Y_i = \gamma_0 + \gamma_1 S_i^2 + \eta_i$.
2. Two separate linear regressions: $\mathcal{T}_i = \beta_0 + \beta_1 S_i + \varepsilon_i^{\mathcal{T}}$ and $\Lambda_i = \delta_0 + \delta_1 S_i + \varepsilon_i^{\Lambda}$.
3. Single generalized regression forest of Y_i on S_i.
4. Two separate generalized regression forests using S_i as the independent variable: one for $\mathcal{T}_i$ and another for Λ_i.

Once we construct those models, we apply them to the same set of 10,000 test customers to evaluate the performance. In particular, we consider the bias (i.e., the average of $\frac{1}{100}\sum_{d=1}^{100}(\widehat{Y}_i^d - S_i^2)$ across 10,000 customers) and the variance (i.e., the average of $\frac{1}{99}\sum_{d=1}^{100}(\widehat{Y}_i^d - S_i^2)^2$ across 10,000 customers) of different approaches.

Figure A.1 showcases the bias and variance of each method across varying noise levels, with $\sigma_{\mathcal{T}}^2$ and σ_{Λ}^2 ranging from 1 to 5. The results indicate that while random effects do not influence a model's bias (given that the bias remains consistent across different noise levels), they profoundly impact the variance of predictions. Notably, the variance associated with the single imputation approach grows exponentially with respect to $\sigma_{\mathcal{T}}^2$ and σ_{Λ}^2. In contrast, this growth rate is considerably more tempered for the separate imputation method. These observations underscore that the separate imputation method can effectively diminish the variance of predicted values by approximately 20% to 80% in comparison to the single imputation approach.

A.3 Further Details about the Simulation Analyses

In this appendix, we provide details about the simulation analyses in Section 1.5 of the main document.



Note: Each dot illustrates the average bias and variance across 10,000 test customers.

Figure A.1: *Bias and Variance Analysis.*

A.3.1 Simulation Setting

For the simulation, we consider a company that conducts a marketing intervention and aims to maximize the total purchase counts ($Y_{i,10}$) over a ten-week time-frame following the intervention. We simulate the data based on the following data generating process:

1. There are two pre-treatment covariates which are drawn i.i.d. from the standard normal distribution, i.e., $X_{i,1}, X_{i,2} \sim_{i.i.d.} \mathcal{N}(0,1)$.

2. At the end of each period, a customer churns with probability $p_{i,t}$.

3. The realized purchase counts in each period when a customer is alive follows a Poisson distribution with mean purchase rate $\lambda_{i,t}$, i.e., $\widetilde{S}_{i,t} \sim \text{Poisson}(\lambda_{i,t})$.

4. The intervention reduces customers' churn probability ($p_{i,t}$) in the *first three periods*. In

particular, the churn probability is

$$\text{(Treatment)} \quad p_{i,t}(W_i = 1) = \begin{cases} \frac{1}{\exp(1.5+0.5X_{i,1}+0.4X_{i,2})}, & \text{if } t \leqslant 3, \\ \frac{1}{\exp(1.4+0.5X_{i,1}+0.4X_{i,2})}, & \text{if } t > 3. \end{cases}$$

$$\text{(Control)} \quad p_{i,t}(W_i = 0) = \frac{1}{\exp(1.4 + 0.5X_{i,1} + 0.4X_{i,2})}, \quad \forall t = 1, \cdots, 10.$$

5. The intervention increases customers' purchase rates in the *first three periods* and has no impact on later periods. The purchase rate follows:

$$\text{(Treatment)} \quad \lambda_{i,t}(W_i = 1) = \begin{cases} \exp(1 + 0.5X_{i,1} + 0.5X_{i,2}), & \text{if } t \leqslant 3, \\ \exp(0.9 + 0.5X_{i,1} + 0.4X_{i,2}), & \text{if } t > 3. \end{cases}$$

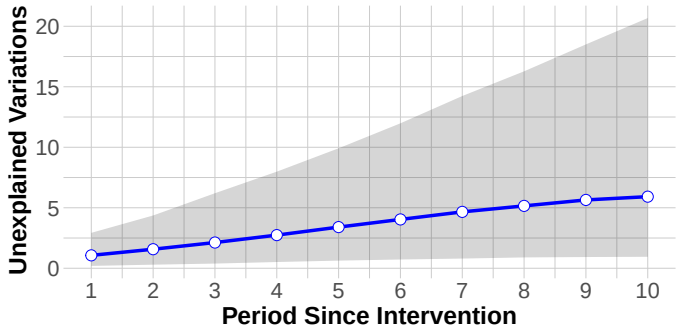
$$\text{(Control)} \quad \lambda_{i,t}(W_i = 0) = \exp(0.9 + 0.5X_{i,1} + 0.4X_{i,2}), \quad \forall t = 1, \cdots, 10.$$

Note that we choose the logit link function for the churn probability to ensure that it lies between 0 and 1. Additionally, we use the exponential link function for the purchase rate to ensure that it is non-negative.

A.3.2 Noise Accumulation Behaviors

We present evidence of noise accumulation due to customer attrition. To do this, we first calculate the unexplained variations as $|Y_{i,T} - \mathbb{E}[Y_{i,T}|\mathbf{X}_i, W_i]|$, where $\mathbb{E}[Y_{i,T}|\mathbf{X}_i, W_i]$ is the expected T-period under the above data generating process. Figure A.2 depicts these unexplained variations in purchases over T periods, ranging from $T = 1$ through $T = 10$. Similar to Figure 1.5, it clearly shows that as the observation period (T) lengthens, the extent of unexplained variations also grows.

Next, we show that unexplained variations have positive serial correlations over different periods. We first calculate the unexplained variations in $S_{i,t}$, i.e., $\varepsilon^S_{i,t} = S_{i,t} - \mathbb{E}[S_{i,t}|\mathbf{X}_i, W_i]$, for $t = 1, \cdots, 10$. Then, we examine the cross-correlation between residuals, given by $\text{Cor}(\varepsilon^S_{i,t_1}, \varepsilon^S_{i,t_2})$. As presented in Figure A.3, there is a persistent positive correlation between per-period unexplained variations for each $1 < t_1 < t_2 \leqslant 10$. Note that our DGP assumes no customer churn in the



Note: Each dot illustrates the median of unexplained variations for all customers in the experimental data. The grey shaded area presents the range between the highest 10% and the lowest 10% of these variations.

Figure A.2: *Unexplained Variations in $Y_{i,t}$ for $T = 1, \cdots, 10$.*

first period, so the unexplained variation $\varepsilon^S_{i,1}$ is expected to be zero as it is purely driven by independent noises in purchase intensity. Therefore, we see zero correlation between $\varepsilon^S_{i,1}$ and $\varepsilon^S_{i,t}$ for $t > 1$.

Note that our DGP assumes no customer churn in the first period, so the unexplained variation $\varepsilon^S_{i,1}$ is expected to be zero as it is purely driven by independent noises in purchase intensity. Therefore, we see zero correlation between between $\varepsilon^S_{i,1}$ and $\varepsilon^S_{i,t}$ for $t > 1$.

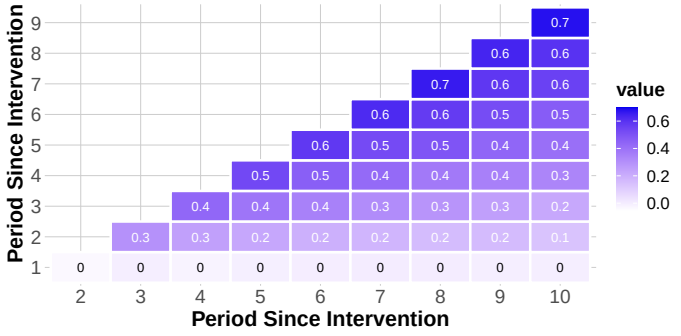


Figure A.3: *Cross-correlation Matrix of Unexplained Variations in Each Period.*

A.3.3 Evaluation Procedure

The following procedure is performed to evaluate the performance of different approaches:

1. Derive the outcome variable $\ddot{Y}$ using the actual outcomes for the default and myopic approaches, or the historical data for the proposed, single, and BG/NBD imputation

approaches.

2. Generate one training set (with $N/2$ treated customers and $N/2$ non-treated customers) and one validation set (with $5,000$ customers for each condition).

3. Construct CATE models ($\widehat{\tau}_{\ddot{Y}}$) using the training set.

4. Calculate the AUTOC value of $\widehat{\tau}_{\ddot{Y}}$ using the validation set.

We generate 200 bootstrap samples and report the mean and standard deviations of each quantity for performance evaluation.

A.3.4 Specification of Surrogate Indices

As discussed in Section 1.5 of the main document, we utilize historical data ($\mathcal{H}$) to generate surrogate indices. Here, we present our model specifications for different imputation methods:

- *(Separate Imputation)* We constructed two linear regressions to predict the observed last purchase time and average purchase rate per active period:

$$\mathcal{T}_{i,10} = \alpha_0^{\mathcal{T}} + \beta_1^{\mathcal{T}} X_{i,1} + \beta_2^{\mathcal{T}} X_{i,2} + \beta_3^{\mathcal{T}} X_{i,1} \cdot X_{i,2} + \sum_{t=1}^{T_0} \left(\gamma_t^{\mathcal{T}} S_{i,t} + \xi_t^{\mathcal{T}} X_{i,1} S_{i,t} + \eta_t^{\mathcal{T}} X_{i,1} S_{i,t} \right) + \varepsilon_i^{\mathcal{T}},$$

$$\Lambda_{i,10} = \alpha_0^{\Lambda} + \beta_1^{\Lambda} X_{i,1} + \beta_2^{\Lambda} X_{i,2} + \beta_3^{\Lambda} X_{i,1} X_{i,2} + \sum_{t=1}^{T_0} \left(\gamma_t^{\Lambda} S_{i,t} + \xi_t^{\Lambda} X_{i,1} S_{i,t} + \eta_t^{\Lambda} X_{i,2} S_{i,t} \right) + \varepsilon_i^{\Lambda},$$

  where $\mathcal{T}_{i,10}$ is the observed last transaction time and $\Lambda_{i,10}$ denotes the average per-period purchase counts until the observed last transaction.

- *(Single Imputation)* We fit the following linear regression to predict $Y_{i,T}$:

$$\begin{aligned} Y_{i,10} = \alpha_0^{Y} + \beta_1^{Y} X_{i,1} + \beta_2^{Y} X_{i,2} + \beta_3^{Y} X_{i,1} X_{i,2} + \\ \sum_{t=1}^{T_0} \left(\gamma_t^{Y} S_{i,t} + \xi_t^{Y} X_{i,1} S_{i,t} + \eta_t^{Y} X_{i,2} S_{i,t} \right) + \varepsilon_i^{Y}. \end{aligned}$$

- *(BG/NBD)* We employ a BG/NBD model with time-invariant covariates $X_{i,1}$ and $X_{i,2}$. Linear specifications were used for all key parameters. After generating expected future purchase counts after T_0, we add it with the observed purchase counts until T_0 to capture the short-term treatment effects.

A.3.5 Specification of CATE Models

In this simulation, we use the following CATE models to corroborate that our findings are not driven by a particular method for CATE estimation:

1. *Causal Forest* [199]: We construct a Causal Forest model using the `grf` package with the default parameters.

2. *R-lasso* [155]: We estimate an R-learner using the `rlearner` package, with both the nuisance models and the CATE function estimated using lasso regression (polynomials of covariates of degree three) with ten-fold cross-validation.

3. *S-GRF*: We predict the conditional expectation of the long-term outcome $\ddot{Y}_i$ given the treatment assignment W_i and the covariates $\mathbf{X}_i$ using a generalized random forest (GRF) model. We then compute the predicted CATE for each validation customer j as the difference between the predicted value given $W_j = 1$ and he predicted value given $W_j = 0$. We implement the GRF model using the `grf` package with the default parameters. [2]

4. *T-GRF*: we construct two GRF models of $\ddot{Y}_i$ on $\mathbf{X}_i$, one for treated customers and another for non-treated customers. We use the default parameters in the `grf` package.

A.3.6 Replication Results of AUTOCs for Different CATE Models

Table A.1 presents the mean and standard deviation of AUTOC values for various approaches using the CATE model as either (i) S-GRF or (ii) T-GRF. These findings align with those from the Causal Forest in Table A.1. Specifically, the separate approach consistently excels over other methods regardless of the CATE estimation model or the size of the training set. In contrast, the single and BG/NBD imputations fall short when compared to the separate and myopic strategies. However, it's noteworthy that all short-term proxies yield superior targeting performance in comparison to the default method, irrespective of the sample size.

[2] We chose the default parameters in the simulation because automatic hyperparameter tuning procedures are time-consuming when the sample size is large (i.e., when $N = 50,000$). However, the results for the $N = 1,000$ case suggest that the findings are almost the same.

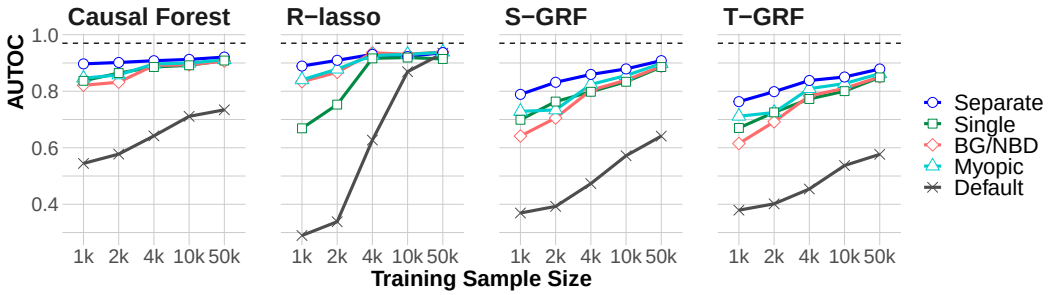
Table A.1: *Comparison of AUTOC Values for Different Outcomes and CATE Models.*

Outcome $\ddot{Y}$	$N = 1,000$		$N = 50,000$	
	S-GRF	**T-GRF**	**S-GRF**	**T-GRF**
Separate Imputation	**0.79 (0.10)**	**0.76 (0.10)**	**0.91 (0.02)**	**0.88 (0.02)**
Single Imputation	0.70 (0.17)	0.65 (0.17)	0.89 (0.02)	0.85 (0.02)
BG/NBD Imputation	0.70 (0.15)	0.68 (0.14)	0.89 (0.02)	0.85 (0.02)
Myopic ($Y_{i,3}$)	0.73 (0.13)	0.71 (0.14)	0.90 (0.02)	0.86 (0.02)
Default ($Y_{i,10}$)	0.40 (0.25)	0.41 (0.25)	0.64 (0.05)	0.58 (0.04)

Higher AUTOC reflects better prioritization rule and therefore superior targeting performance. We average the results over 200 replications and show in parentheses the standard deviation.

A.3.7 Sample Size Efficiency

We investigate the sample size efficiency of different approaches. Figure A.4 displays the AUTOC values for various CATE models and training sample sizes. The results indicate that CATE models utilizing short-term proxies consistently outperform the default approach, regardless of the training sample size. Moreover, the separate imputation method persistently surpasses other methods in terms of AUTOC for the same CATE model and training sample size. These findings suggest that incorporating short-term outcomes is a viable strategy for enhancing targeting performance, as it enables more accurate CATE estimation without requiring substantially larger sample sizes.

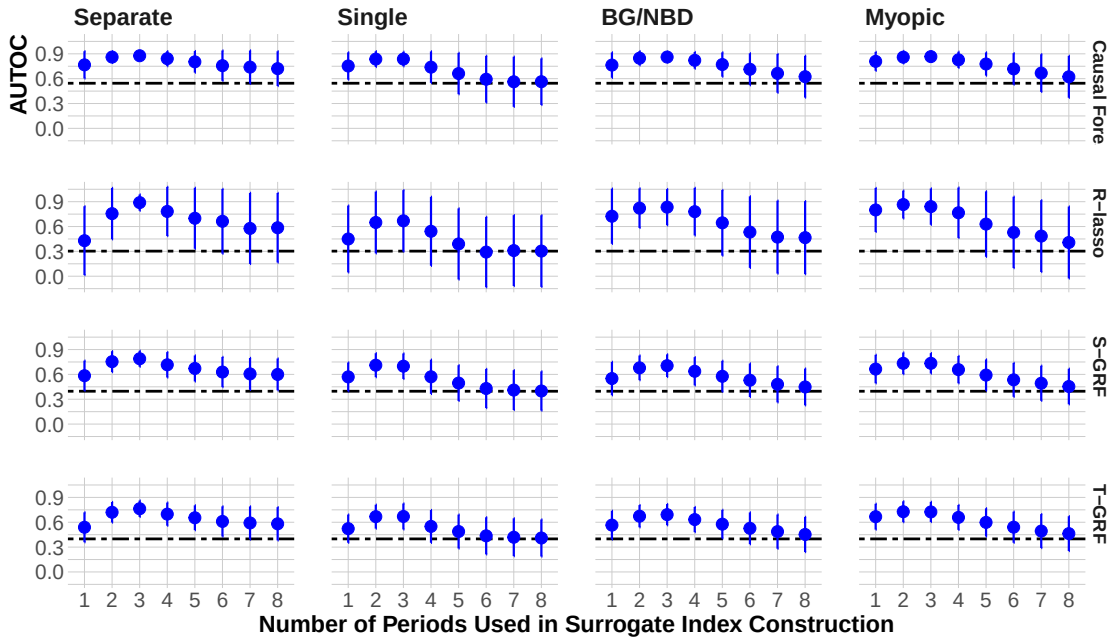


Note: Each point reports the average over 200 simulation replications. The dashline represents the AUTOC value when the prioritization rule is based on the true CATE.

Figure A.4: *Area-under-TOC Curves: CATE Models with Different Training Sample Size.*

A.3.8 Trade-off between Information and Noise Accumulation

We reproduce the analyses described in Section 1.5.5 for different approaches and CATE models. Figure A.5 shows the results of AUTOCs for various CATE models and the number of periods used for surrogacy construction. Our results indicate that (i) using short-term proxies consistently leads to higher AUTOC compared to using the actual long-term outcome (except for R-lasso), (ii) the separate imputation method outperforms other short-term proxies regardless of the CATE models used for estimation, (iii) the separate imputation method shows higher robustness to noise, as the AUTOC declines more slowly, and (iv) using $T_0 = 3$ (i.e., the minimal number of periods such that the surrogacy assumption is satisfied) generally yields the most effective targeting performance.



Note: Each point reports the average over 200 simulation replications together with the two standard error interval. The larger the AUTOC, the better targeting performance.

Figure A.5: *Area-under-TOC Curve: CATE Models for Outcomes Using Different Periods of Information.*

A.4 Further Details for the Empirical Application

In this appendix, we provide additional analyses for the empirical application presented in Section 1.6 of the main document.

A.4.1 Specification of Surrogate Indices

To construct the surrogate indices, we gathered historical data of customers who were acquired at least ten weeks before the experiment started (4,031 in total) and imputed different outcome variables as follows:

- $\widetilde{Y}_{10}^{\text{Single}}(S_{i,1}, \mathbf{X}_i)$: we fit a random forest model [15] of $Y_{i,10}$ on the first-week purchase ($S_{i,1}$) and customer covariates ($\mathbf{X}_i$), reported in Appendix 1.6.1. We perform automatic parameter tuning using the function provided by the `grf` package.

- $\widetilde{Y}_{10}^{\text{Sep}}(S_{i,1}, \mathbf{X}_i)$: we fit two random forests, one for the observed last purchase week ($\mathcal{T}_{i,10}$) and another for the purchase counts per period until the last purchase week ($\Lambda_{i,10} \equiv Y_{i,10}/\mathcal{T}_{i,10}$), on $S_{i,1}$ and $\mathbf{X}_i$. We perform automatic parameter tuning using the function provided by the `grf` package.

- $\widetilde{Y}_{10}^{\text{BTYD}}(S_{i,1}, \mathbf{X}_i)$: we fit a BG/NBD model with $\mathbf{X}_i$ as time-invariant covariates.

A.4.2 Specification of CATE Models

Given an outcome variable $\ddot{Y}_i$, we construct four types of CATE models to show robustness of our findings, including:

1. *S-learner*: we predict $\mathbb{E}[\ddot{Y}_i|W_i, \mathbf{X}_i]$ by regressing $\ddot{Y}_i$ on $W_i, \mathbf{X}_i$ using random forest and perform the automatic hyperparameter tuning using the method implemented in the `grf` package.

2. *T-learner*: we construct two random forests of $\ddot{Y}_i$ on $\mathbf{X}_i$, one for treated customers and another for non-treated customers. We perform the automatic hyperparameter tuning using the method implemented in the `grf` package.

3. *X-learner* [134]: all the outcome models in X-learner are estimated using the random forest with automatic hyperparameter tuning. We estimate the propensity score using the probability forest implemented in the `grf` package.

4. *Causal Forest* [199]: we use the causal forest function implemented in the `grf` package with automatic hyperparameter tuning.

A.4.3 Details for Policy Learning Using Doubly Robust Scores

In this section, we provide a detailed explanation of how we implement doubly robust policy learning, as proposed by [17]. Specifically, for each training-validation split, we learn the policy by the following steps:

1. Compute the outcome variable $\ddot{Y}$ for the training set.

2. Compute the (honest) doubly robust score for i in the training set:

$$\widehat{\Gamma}_i = \left[\widehat{m}^1(\mathbf{X}_i) - \widehat{m}^0(\mathbf{X}_i)\right] + \frac{W_i - \widehat{e}(\mathbf{X}_i)}{\widehat{e}(\mathbf{X}_i)}\left[Y_i - \widehat{m}^{W_i}(\mathbf{X}_i)\right].$$

   We use `grf` and `policytree` packages [188] to derive the doubly robust score for each customer.

3. Derive the targeting policy $\widehat{\pi} : \mathbf{X}_i \to \{0, 1\}$ by solving the optimization problem:

$$\widehat{\pi} = \arg\max_{\pi} \sum_{i \in \mathcal{D}_{\text{train}}} [2\pi(\mathbf{X}_i) - 1]\,\widehat{\Gamma}_i,$$

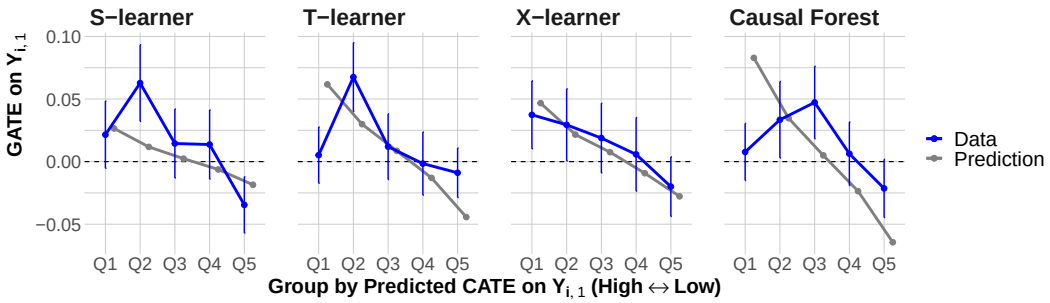
   where we constrain π in the class of probability forest in the `grf` package.

A.4.4 Replication Results for the Motivating Example

Targeting for Short-term Outcome.

In the motivating example (Section 1.2 of the main document), we show that the focal company can develop an effective CATE model when the outcome variable is $Y_{i,1}$. Figure A.6 presents the GATEs on $Y_{i,1}$ across quintiles based on predicted CATE. This chart is constructed similarly to Figure 1.2, now presenting the results for the different CATE models. While the actual CATE curves

are not perfectly decreasing for all models, they are still effective in distinguishing customers with high CATEs from those with low CATEs, as $\mathcal{Q}_1, \mathcal{Q}_2, \mathcal{Q}_3$ have higher treatment effects than $\mathcal{Q}_4$ and $\mathcal{Q}_5$ have.

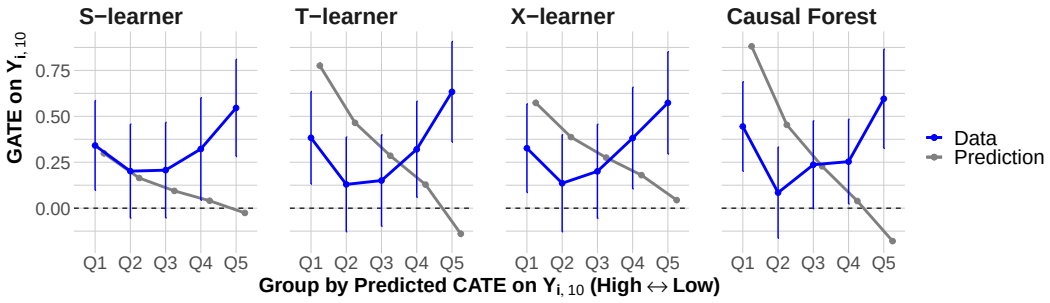


Note: Each point represents the mean of bootstrap results on the validation customers together with the one standard deviation interval. Groups $\mathcal{Q}_1, \cdots, \mathcal{Q}_5$ are categorized based on the decreasing order of the predicted CATE.

Figure A.6: *CATE Models for* $Y_{i,1}$.

CATE Models for Long-term Outcome.

Similarly, Figure A.7 illustrates the predicted and actual GATEs when using $Y_{i,10}$ as outcome variable. Notably, all CATE models produce the same V-shaped curve, indicating that the firm would overlook a significant proportion (e.g., the bottom quintile group $\mathcal{Q}_5$) of the "should-target" customers when the targeting policy is based on these models.

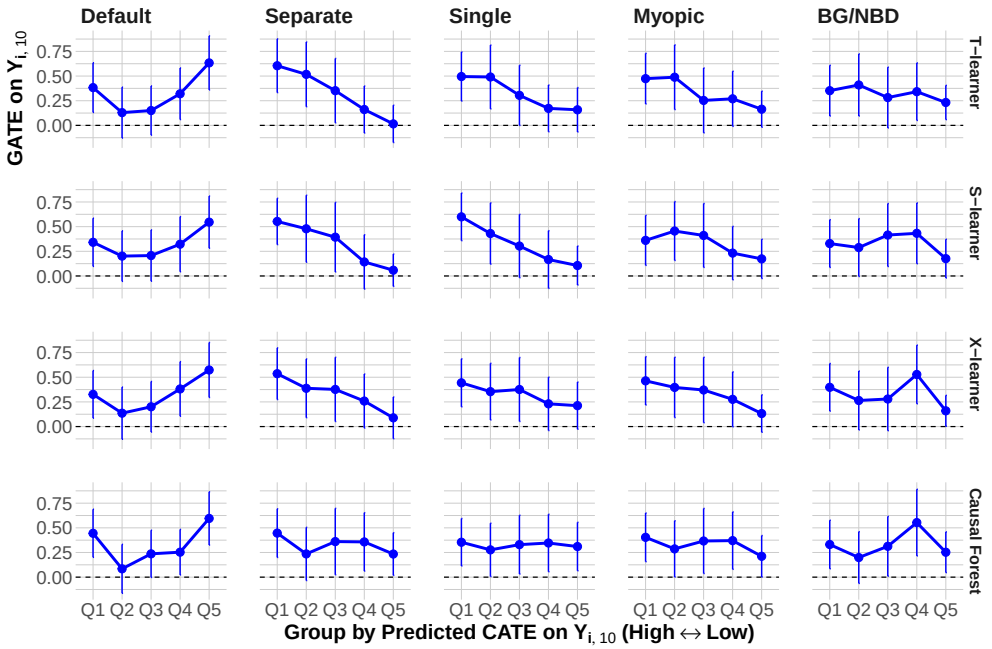


Note: Each point represents the mean of bootstrap results on the validation customers together with the one standard deviation interval. Groups $\mathcal{Q}_1, \cdots, \mathcal{Q}_5$ are categorized based on the decreasing order of the predicted CATE.

Figure A.7: *CATE Models for* $Y_{i,10}$.

A.4.5 Replication Results for the GATE Analysis

Figure A.8 displays the GATEs by predicted CATE levels of different outcome variables. Notably, regardless of the methods used, the separate imputation method consistently produces the steepest CATE curve, suggesting the superiority of targeting based on our proposed solution. Note that both Single and Myopic approaches also result in reasonably good performance when compared to models based on $Y_{i,10}$.



Note: Each point represents the mean of bootstrap results on the validation customers together with the one standard deviation interval (the errorbar). Groups $\mathcal{Q}_1^{\ddot{Y}}, \cdots, \mathcal{Q}_5^{\ddot{Y}}$ are defined by decreasing order of treatment effect, as predicted by the CATE model for different outcome variables. GATEs are computed from $Y_{i,10}$.

Figure A.8: *Actual group average treatment effects by predicted CATE levels on different outcome variables.*

A.4.6 Replication Results for Policy Improvement

In this appendix, we replicate the expected profit improvement results in Table 1.3 of the main document using the CATE models outlined in Appendix A.4.2. The results demonstrate the robustness of our findings: regardless of the CATE model used for estimation, the proposed separate imputation approach consistently yields the best profitability. Additionally, all targeting policies based on short-term outcomes consistently yield a positive profit improvement, while

targeting using the default approach results in profit loss.

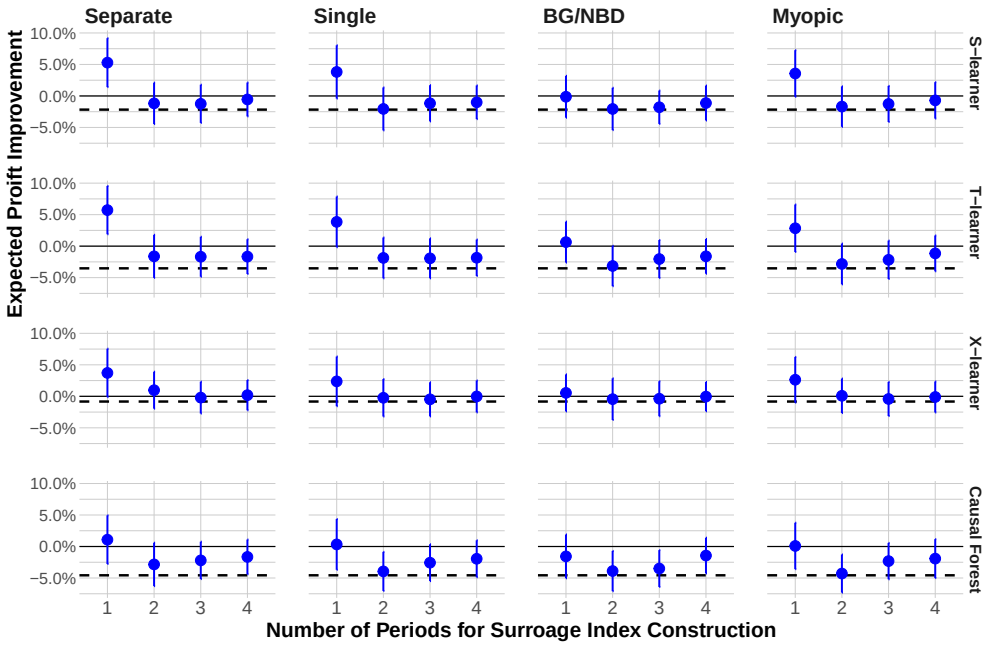
Table A.2: *Expected Profit Improvement: Targeting Based on Different Models.*

Outcome Variable	T-learner	S-learner	X-learner	Causal Forest	Policy Learning
Separate Imputation	5.81% (4.19%)	5.66% (4.27%)	3.98% (3.53%)	1.64% (4.32%)	4.57% (3.90%)
Single Imputation	4.06% (4.51%)	4.45% (5.17%)	2.29% (4.31%)	−1.44% (4.71%)	3.66% (4.75%)
BG/NBD Imputation	0.95% (2.77%)	1.67% (2.80%)	1.93% (2.32%)	1.14% (4.33%)	−0.26% (4.36%)
Myopic	3.34% (3.60%)	3.08% (3.65%)	2.96% (2.94%)	1.09% (4.41%)	3.51% (4.58%)
Default	−3.52% (3.17%)	−1.75% (2.86%)	−0.89% (2.85%)	−4.21% (3.7-%)	−3.44% (3.37%)

We average the profit improvement over 200 replications and show in parentheses the standard deviations.

A.4.7 How Many Periods Should the Focal Firm Use?

Figure A.9 compares the expected profit improvement from targeting based on surrogate indices constructed using different periods of outcomes. The result suggests that using one-period outcome in surrogate models gives the highest profit, and targeting based on short-term signals consistently outperforms or is as good as targeting based on the actual long-term outcome.



Note: Each point represents the mean of bootstrap results on the validation customers together with the one standard deviation interval (the errorbar). The dashed line reports the expected profit improvement when the targeting policy is based on predicted CATEs on $Y_{i,10}$.

Figure A.9: *Expected Profit Improvement: Comparison of Different Periods Used for Surrogate Models.*

Appendix B

Appendix to Chapter 2

B.1 Additional Literature Review

B.1.1 Classical Measurement Error Bias Correction

This section reviews existing approaches for correcting classical measurement error bias, summarized in Table B.1.

1. *Repeated Measurements of Variables*: These techniques aim to mitigate random noise by averaging multiple measurements of the same variable, reducing the impact of individual errors. While they are theoretically capable of handling complex functional forms, they become impractical in high-dimensional settings due to the exponential increase in the number of repeated measurements required as the dimensionality of the data grows. This requirement can be resource-intensive or even infeasible in practice. Additionally, these techniques pose significant privacy risks, as repeated data collection increases the likelihood of re-identifying individuals. This undermines privacy protections, making such methods unsuitable for use in contexts requiring strict privacy constraints.

2. *Additional Clean Data*: These methods calibrate the model using a separate dataset with noise-free measurements to correct estimates from the larger dataset. It is theoretically well-suited in high-dimensional settings with flexible functional forms and model-agnostic applicability. However, it raises significant privacy concerns because the clean dataset itself may contain sensitive information.

3. *Instrumental Variables* (IVs): These techniques use additional instrumental variables (IVs) that are correlated with the noisy covariate but uncorrelated with the measurement error or outcome, serving as proxies to correct bias. However, identifying valid instruments in high-dimensional settings is challenging, as suitable instruments are needed for many covariates. Additionally, IV approaches often rely on linear or parametric assumptions, limiting their ability to capture heterogeneity. Their integration with specific models also reduces model-agnostic flexibility. Moreover, protecting instrumental variables under LDP can exacerbate the weak instruments problem, further complicating the estimation process.

4. *Noise Distribution Information*: These methods use techniques such as deconvolution kernel estimators or the generalized method of moments (GMM) to correct bias by leveraging information about the noise distribution. However, these approaches are not well-suited to our context for several reasons. First, while effective in low-dimensional settings, these methods face challenges in high-dimensional contexts due to data sparsity, which limits kernel methods, and the increasing number of moment conditions, which complicates GMM. Second, they typically assume fixed functional forms, such as linear models, limiting flexibility. Deconvolution techniques depend on specific smoothing processes and often assume normal noise, while GMM requires explicitly derived moment conditions, making it impractical for nonparametric models. Third, neither method is model-agnostic or capable of dynamically adjusting to covariate-specific bias patterns, as they lack mechanisms for localized corrections.

5. *Higher-order Moments*: These methods utilize the overidentifying information in higher-order moments of the data to correct measurement error bias, assuming the measurement errors satisfy certain independence conditions. While this approach can help recover unbiased estimates without requiring external instruments, its applicability is limited to relatively simple cases, such as linear relationships or settings with a single covariate.

B.1.2 Comparison of Proposed Approach to Existing Sample Splitting Approaches.

Table B.2 summarizes the distinctions between our proposed method and the sample-splitting strategies employed in causal forests [14, 199] and transformed outcome regressions for CATE

Table B.1: *Comprehensive Overview of Solutions for Measurement Error Bias*

Study	Flexible with Many Covariates	Model Specification	Estimation Method
Repeated Measurements of Variables			
Hausman et al. (1995)	Yes	Polynomial	Least Square
Li (2002)	Yes	Nonlinear	Least Square
Schennach (2004)	Yes	Polynomial	GMM
Agarwal & Singh (2024)	Yes	Linear	PCA + Least Square
Additional Clean Data			
Bound et al. (1989)	Yes	Linear	OLS
Sepanski & Carroll (1993)	No	Nonlinear	OLS
Chen et al. (2005)	Yes	Nonlinear	GMM
Hu & Ridder (2012)	No	Nonlinear	Deconvolution + MLE
Instrumental Variables			
Hausman et al. (1991)	No	Polynomial	GMM
Newy (2001)	No	Nonlinear	GMM
Schennach (2007)	No	Nonlinear	GMM
Hu & Schennach (2008)	No	Nonlinear	MLE
Information of Noise Distribution			
Wolter & Fuller (1982)	No	Polynomial	Parametric MLE
Fan & Truong (1993)	No	Nonlinear	Deconvolution + Kernel Estimator
Higher-order Moments in Data			
Manoranjan (1980)	No	Linear	GMM
Erickson & Whited (2002)	Yes	Linear	GMM
Bonhomme & Robin (2010)	No	Linear	GMM
Schennach & Hu (2013)	No	Known functional form	MLE

estimation [41, 155, 126, 206]. In causal trees/forests, sample splitting is primarily used to decouple the *partitioning* procedure from the *leaf-level treatment-effect estimation*, preventing the tree from overfitting the same data used for effect estimation. While this approach is in the same spirit of reducing bias through data partitioning, our goal focuses on avoiding overfitting when *calibrating* an already-fitted CATE model using a noisy CATE proxy — noise that arises naturally under local differential privacy (LDP).

In standard transformed outcome regressions, the focus is on ensuring that the nuisance models (e.g., outcome regressions and propensity scores) are trained on observations disjoint from those used to construct the proxy CATE scores. This separation prevents bias in the nuisance models from influencing the CATE model derived from the proxy scores. Our framework goes a step further by splitting the dataset for CATE estimation into three disjoint subsets: one for constructing the initial CATE model, another for model calibration, and a third for model validation. This additional splitting ensures that the calibration process is entirely independent of both the initial model construction and the validation phase. Moreover, during boosting, our

approach ensures that the construction of the calibration model is independent of the step size determination. This separation introduces a novel element not found in existing boosting-based CATE models to enhance model robustness and reduce the risk of overfitting.

Table B.2: *Key Distinctions Between Our Method and Existing Sample-splitting Approaches*

Method	Where Does Splitting Occur?	Primary Goal
Causal Tree / Forest [14, 16, e.g.,]	Partition tree structure vs. Estimate leaf-level treatment effect.	Decouple treatment effect estimation from tree partitioning
Transformed Outcome Regression [155, 180, 126, 206]	Estimate nuisance models vs. Construct the proxy score and CATE models.	Eliminate bias from nuisance estimates
Our Proposed Method	1. Train the initial CATE model vs. Calibrate the initial model vs. Validate the calibration process. 2. Estimate calibration model vs. Determine the optimal step size.	Avoid overfitting to the noisy proxy at every step of boosting

B.2 Theoretical Analysis of LDP Impact on Bias and Variance

In Section 2.3, we provide intuition on how LDP protection impacts bias and variance. We now present a formal analysis of bias and variance in widely used CATE models. Specifically, we consider the class of CATE estimators studied in [112].

Assumption B.1 ***(Assumption in App-2 of [112])*** *For a given individual with covariates $\mathbf{x}_{new}$, the predicted CATE $\widehat{\tau}(\mathbf{x}_{new} \mid \mathcal{D}^o, \mathcal{D}^\ell, \mathcal{D}^m)$ can be expressed as the difference between two (adjusted) outcome estimators, $\widehat{\mu}^0$ and $\widehat{\mu}^1$, in the following form:*

$$\widehat{\mu}^w(\mathbf{x}_{new}) = \sum_{i \in \mathcal{I}^o : W_i = w} \widehat{\ell}_i^w(\mathbf{x}_{new} \mid \mathcal{D}^\ell)[Y_i - \widehat{m}^w(\mathbf{X}_i \mid \mathcal{D}^m)],$$

where $\mathcal{I}^o$ denotes the set of individuals used to impute the two outcome predictions, $\mathcal{D}^\ell = \{\mathcal{D}^\ell, \mathcal{Y}^\ell, \mathcal{W}^\ell\}$ represents the data used to construct the weight function $\widehat{\ell}_i^w(\mathbf{x}_{new} \mid \mathcal{D}^\ell)$, and $\mathcal{D}^m = \{\mathcal{D}^m, \mathcal{Y}^m, \mathcal{W}^m\}$ is the data used to determine the adjustment function $\widehat{m}^w(\mathbf{X}_i \mid \mathcal{D}^m)$. We also denote $\mathcal{D}^o = \{\mathcal{X}^o, \mathcal{Y}^o, \mathcal{W}^o\}$ as the experiment data of individuals in $\mathcal{I}_o$, Furthermore, we assume that the estimation process of the CATE model satisfies the following conditions:

1. *(Honest Estimation) The weight function is independent of Y_j, $\forall j \in \mathcal{D}^o$. In other words, $\mathcal{D}^\ell$ is either independent of $\mathcal{D}^o$ or depends only on the covariate values of individuals in $\mathcal{I}^o$.*

2. *(Cross Fitting) The adjustment function is either zero or constructed from the dataset $\mathcal{D}^m$ that is independent of both $\mathcal{D}^o$ and $\mathcal{D}^\ell$.*

In Proposition App-1 of [112], it is shown that this class of CATE models includes a wide range of approaches, including S-learner, T-learner, Causal Forest, and R-learner, when combined with nonlinear machine learning methods such as random forests and kernel regressions. For neural network models, while explicitly writing the weight function in closed form is challenging, prior research has shown that neural network models can be interpreted as a form of kernel regression (see Golikov, 2020, for a survey). Specifically, the kernel function in this case corresponds to the neural tangent kernel (Jacot et al., 2018). Since our model class already includes kernel regression methods, it should also be applicable to neural network models. For double machine learning methods, such as those proposed by **(author?)** [180, 126], the key distinction from R-learners is that they use the AIPW score instead of Robinson's transformation to estimate the proxy CATE. Additionally, they apply separate adjustment functions for the treatment and control groups rather than a single adjustment function for all individuals. Given this structure, we can easily see that the double machine learning methods also fall within the model class considered in our analyses.

B.2.1 Bias Analysis

To apply the Taylor approximation approach in **(author?)** [43] to characterize the bias of the above class of the CATE model, we further impose the following smoothness assumptions on the CATE models:

Assumption B.2 ***(Smoothness Conditions)***

1. *(Smoothness) The weight function $\widehat{\ell}_i^w(\cdot \mid \mathcal{D}^\ell)$, and the adjustment function, $\widehat{m}^w(\cdot \mid \mathcal{D}^\ell)$ are C^∞-smooth, i.e., they have derivatives of all orders .*[1]

2. *(Finite Variations) There exists a constant K such that the k-th order derivatives of $\widehat{\ell}_i^w(\cdot \mid \mathcal{D}^\ell)$ and the adjustment function $\widehat{m}^w(\cdot \mid \mathcal{D}^\ell)$ exist for all $k \geqslant K$. (Note: If this assumption is not imposed, the bias and variance introduced by LDP protection could become unbounded.)*

[1] While the weight function is not differentiable for tree-based methods, we can use the differentiable *smooth bump function* [79] to approximate the weight function. Therefore, the Taylor approximation of the bias is still applicable for tree-based methods.

3. *(Bounded Sensitivity) The sensitivity of the weight function towards a specific covariate of each individual is bounded by a dominating function with a finite integral. That is, for any customer i and covariate p, we have*

$$\left| \frac{\widehat{\ell}_i^w(x_{i,p} + \delta_{i,p}, \mathbf{x}_{-(i,p)} \mid \mathcal{D}^\ell) - \widehat{\ell}_i^w(x_{i,p}, \mathbf{x}_{-(i,p)} \mid \mathcal{D}^\ell)}{\delta_{i,p}} \right| < g(x_{i,p}, \mathbf{x}_{-(i,p)})$$

for any value of $\mathbf{x}_{-(i,p)}$ and bounded $\delta_{i,p}$, where $\int_{-\infty}^{\infty} g(x_{i,p}, \mathbf{x}_{-(i,p)}) d\mathbf{x}_{-(i,p)} < \infty$. Similarly, we assume that the the adjustment function $\widehat{m}$ satisfies the bounded sensitivity assumption. This condition ensures that the derivative and expectation operators are exchangeable.

Note that the smoothness assumption is introduced to ensure the existence of derivatives and the validity of the Taylor approximation theorem. In fact, many nonlinear machine learning models, such as kernel regressions or neural networks with sigmoid activation functions, are inherently C^∞ smooth. The finite-variation assumption is imposed to ensure that the bias remains bounded. Without this assumption, the bias could be expressed as a weighted average of infinitely many higher-order moments, where the weights correspond to the higher-order derivatives of the CATE model. The third and fourth assumptions are standard for the Talyor approximation theorem.

We now formally state the Taylor approximation theorem for the bias in the following result and provide its proof.

Theorem B.1 ***(Bias Analysis)*** *Assume that the CATE model $\widehat{\tau}$ belongs to the model class in Assumption B.1 satisfies the smoothness conditions in Assumption B.2. Let $\mathcal{D} = \{\mathcal{D}^\ell, \mathcal{D}^m, \mathcal{D}^o\}$ denote the non-private data and $\widetilde{\mathcal{D}} = \{\widetilde{\mathcal{D}}^\ell, \widetilde{\mathcal{D}}^m, \widetilde{\mathcal{D}}^o\}$ denote the LDP-protected data. For simplicity, consider both Y_i and $\mathbf{X}_i$ are perturbed by additive noise from the same distribution with mean zero and a k-th order moment of σ_k. Then, the additional bias in $\widehat{\tau}$ introduced by the LDP protection can be approximated as:*

$$\mathbb{E}\left[\widehat{\tau}(\widetilde{\mathbf{x}}_{new} \mid \widetilde{\mathcal{D}}) - \widehat{\tau}(\mathbf{x}_{new} \mid \mathcal{D})\right] =$$

$$\sum_{i \in \mathcal{I}^o} (2W_i - 1) \left\{ \mathbb{E}\left[Y_i - \widehat{m}^w(\mathbf{X}_i \mid \mathcal{D}^m)\right] \Delta_i^\ell(\mathbf{x}_{new}) - \mathbb{E}\left[\widehat{\ell}_i^w(\mathbf{x}_{new}, \mathcal{D}^\ell)\right] \cdot \Delta^m - \Delta_i^\ell(\mathbf{x}_{new}) \cdot \Delta_i^m \right\},$$

where $\Delta_i^\ell(\mathbf{x}_{new}) = \sum_{k=1}^K \frac{1}{k!} \sigma_k trace\left(\mathbb{E}\left[\partial^k_{\mathbf{x}_{new}, \mathcal{D}^\ell} \widehat{\ell}_i^{W_i}(\mathbf{x}_{new} \mid \mathcal{D}^\ell)\right]\right)$ represents the expected impact of LDP noise on the weight function, and $\Delta_i^m = \sum_{k=1}^K \frac{1}{k!} \sigma_k trace\left(\mathbb{E}\left[\partial^k_{\mathbf{X}_i, \mathcal{D}^m} \widehat{m}^w(\mathbf{X}_i \mid \mathcal{D}^m)\right]\right)$ represents the expected impact of LDP noise on the adjustment function.

Proof. First, we can write the additional bias as

$$\mathbb{E}\left[\widehat{\tau}(\widetilde{\mathbf{x}}_{\text{new}} \mid \widetilde{\mathcal{D}}) - \widehat{\tau}(\mathbf{x}_{\text{new}} \mid \mathcal{D})\right] = \mathbb{E}\left[\widetilde{\widehat{\mu}}^1(\widetilde{\mathbf{x}}_{\text{new}} \mid \widetilde{\mathcal{D}}) - \widehat{\mu}^1(\mathbf{x}_{\text{new}} \mid \mathcal{D})\right] - \mathbb{E}\left[\widetilde{\widehat{\mu}}^0(\widetilde{\mathbf{x}}_{\text{new}} \mid \widetilde{\mathcal{D}}) - \widehat{\mu}^0(\mathbf{x}_{\text{new}} \mid \mathcal{D})\right],$$

where $\widetilde{\widehat{\mu}}^w(\widetilde{\mathbf{x}}_{\text{new}} \mid \widetilde{\mathcal{D}}) = \sum_{i \in \widetilde{\mathcal{D}}^o : W_i = w} \widehat{\ell}_i^w(\widetilde{\mathbf{x}}_{\text{new}} \mid \widetilde{\mathcal{D}}^\ell)[Y_i - \widehat{m}^w(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}^m)]$ is the implied outcome model under LDP protection.

Next, we derive the difference in the implied outcome models, i.e., $\mathbb{E}\left[\widetilde{\widehat{\mu}}^w(\widetilde{\mathbf{x}}_{\text{new}} \mid \widetilde{\mathcal{D}}) - \widehat{\mu}^w(\mathbf{x}_{\text{new}} \mid \mathcal{D})\right]$. We start by considering the difference for each individual i in $\mathcal{D}^o$ with treatment assignment w. We can write the total impact of the injected noises on the predicted CATE can be written as

$$\begin{aligned}
&\widehat{\ell}_i^w(\widetilde{\mathbf{x}}_{\text{new}} \mid \widetilde{\mathcal{D}}^\ell)\left[\widetilde{Y}_i - \widehat{m}^w(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}_m)\right] - \widehat{\ell}_i^w(\mathbf{x}_{\text{new}} \mid \mathcal{D}^\ell)\left[Y_i - \widehat{m}^w(\mathbf{X}_i \mid \mathcal{D}^m)\right] \\
&= \underbrace{\left[\widehat{\ell}_i^w(\widetilde{\mathbf{x}}_{\text{new}} \mid \widetilde{\mathcal{D}}^\ell) - \widehat{\ell}_i^w(\mathbf{x}_{\text{new}} \mid \mathcal{D}^\ell)\right]\left[Y_i - \widehat{m}^w(\mathbf{X}_i \mid \mathcal{D}^m)\right]}_{A_i} - \\
&\underbrace{\left[\widehat{\ell}_i^w(\widetilde{\mathbf{x}}_{\text{new}} \mid \widetilde{\mathcal{D}}^\ell) - \widehat{\ell}_i^w(\mathbf{x}_{\text{new}} \mid \mathcal{D}^\ell)\right]\left[\widehat{m}^w(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}_m) - \widehat{m}^w(\mathbf{X}_i \mid \mathcal{D}^m)\right]}_{B_i} - \\
&\underbrace{\widehat{\ell}_i^w(\mathbf{x}_{\text{new}} \mid \mathcal{D}^\ell)\left[\widehat{m}^w(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}_m) - \widehat{m}^w(\mathbf{X}_i \mid \mathcal{D}^m)\right]}_{C_i} + \underbrace{\widehat{\ell}_i^w(\mathbf{x}_{\text{new}} \mid \mathcal{D}^\ell)\eta_i}_{D_i}.
\end{aligned} \tag{B.1}$$

Note that $\mathbb{E}[D_i] = 0$ since $\mathbb{E}[\eta_i] = 0$ and $\widehat{\ell}_i^w(\mathbf{x}_{\text{new}} \mid \mathcal{D}^\ell)$ is independent of η_i due to honest estimation.

Now, our objective is to determine the approximate discrepancies in the weight and adjustment functions, namely, $\widehat{\ell}_i^w(\widetilde{\mathbf{x}}_{\text{new}} \mid \widetilde{\mathcal{D}}^\ell) - \widehat{\ell}_i^w(\mathbf{x}_{\text{new}} \mid \mathcal{D}^\ell)$ and $\widehat{m}^w(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}_m) - \widehat{m}^w(\mathbf{X}_i \mid \mathcal{D}^m)$. Using Taylor approximation, the approximate deviation in weight functions due to the introduced noises in $\widetilde{\mathbf{x}}_{\text{new}}$ and $\widetilde{\mathcal{X}}^\ell$ can be written as:

$$\widehat{\ell}_i^w(\widetilde{\mathbf{x}}_{\text{new}} \mid \widetilde{\mathcal{D}}^\ell) - \widehat{\ell}_i^w(\mathbf{x}_{\text{new}} \mid \mathcal{D}^\ell) = \sum_{k=1}^{K}\left(\frac{1}{k!}\partial^k_{\mathbf{x}_{\text{new}},\mathcal{D}^\ell}\widehat{\ell}_i^w(\mathbf{x}_{\text{new}} \mid \mathcal{D}^\ell)\right)\boldsymbol{\eta}^{\circ k}_{\mathbf{x}_{\text{new}},\mathcal{D}^\ell},$$

where $\partial^k_{\mathbf{x}_{\text{new}},\mathcal{D}^\ell}\widehat{\ell}_i^w(\mathbf{x}_{\text{new}} \mid \mathcal{D}^\ell)$ denotes the vector of all k-th order directional derivatives of $\widehat{\ell}_i^w$ with respect to $\mathbf{x}_{\text{new}}$ and $\mathcal{D}^\ell$, and $\boldsymbol{\eta}_{\mathbf{x}_{\text{new}},\mathcal{D}^\ell}$ represents the vector of noises being injected into $\mathbf{x}_{\text{new}}$ and $\mathcal{D}^\ell$. Here, $Z^{\circ k}$ denotes a vector or matrix in which each element is raised to the power k from the corresponding element in Z. Similarly, we can write the impact of injected noises on the

adjustment function as:

$$\widehat{m}^w(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}_m) - \widehat{m}^w(\mathbf{X}_i \mid \mathcal{D}^m) = \sum_{k=1}^{K} \left(\frac{1}{k!} \partial^k_{\mathbf{X}_i, \mathcal{D}^m} \widehat{m}^w(\mathbf{X}_i \mid \mathcal{D}^m) \right) \boldsymbol{\eta}^{\circ k}_{\mathbf{X}_i, \mathcal{D}^m}.$$

Since $\mathcal{D}^o$, $\mathcal{D}^\ell$, and $\mathcal{D}^m$ are independent, and the noise injected into the data are i.i.d. and independent from the experiment data, we have

$$\mathbb{E}\left[\widehat{\ell}^w_i(\widetilde{\mathbf{x}}_{\text{new}} \mid \widetilde{\mathcal{D}}^\ell) - \widehat{\ell}^w_i(\mathbf{x}_{\text{new}} \mid \mathcal{D}^\ell)\right] = \sum_{k=1}^{K} \frac{1}{k!} \sigma_k \text{trace}\left(\mathbb{E}\left[\partial^k_{\mathbf{x}_{\text{new}}, \mathcal{D}^\ell} \widehat{\ell}^w_i(\mathbf{x}_{\text{new}} \mid \mathcal{D}^\ell)\right]\right),$$

$$\mathbb{E}\left[\widehat{m}^w(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}_m) - \widehat{m}^w(\mathbf{X}_i \mid \mathcal{D}^m)\right] = \sum_{k=1}^{K} \frac{1}{k!} \sigma_k \text{trace}\left(\mathbb{E}\left[\partial^k_{\mathbf{X}_i, \mathcal{D}^m} \widehat{m}^w(\mathbf{X}_i \mid \mathcal{D}^m)\right]\right).$$

Now, we can write the expectation of A_i in (B.1) as

$$\mathbb{E}[A_i] = \mathbb{E}[Y_i - \widehat{m}^w(\mathbf{X}_i \mid \mathcal{D}^m)] \cdot \sum_{k=1}^{K} \frac{1}{k!} \sigma_k \text{trace}\left(\mathbb{E}\left[\partial^k_{\mathbf{x}_{\text{new}}, \mathcal{D}^\ell} \widehat{\ell}^w_i(\mathbf{x}_{\text{new}} \mid \mathcal{D}^\ell)\right]\right).$$

Similarly, the expectation of B_i and C_i in (B.1) can be written as

$$\mathbb{E}[B_i] = \left(\sum_{k=1}^{K} \frac{1}{k!} \sigma_k \text{trace}\left(\mathbb{E}\left[\partial^k_{\mathbf{x}_{\text{new}}, \mathcal{D}^\ell} \widehat{\ell}^w_i(\mathbf{x}_{\text{new}} \mid \mathcal{D}^\ell)\right]\right)\right) \times \left(\sum_{k=1}^{K} \frac{1}{k!} \sigma_k \text{trace}\left(\mathbb{E}\left[\partial^k_{\mathbf{X}_i, \mathcal{D}^m} \widehat{m}^w(\mathbf{X}_i \mid \mathcal{D}^m)\right]\right)\right),$$

$$\mathbb{E}[C_i] = \mathbb{E}\left[\widehat{\ell}^w_i(\mathbf{x}_{\text{new}}, \mathcal{D}^\ell)\right] \cdot \sum_{k=1}^{K} \frac{1}{k!} \sigma_k \text{trace}\left(\mathbb{E}\left[\partial^k_{\mathbf{X}_i, \mathcal{D}^m} \widehat{m}^w(\mathbf{X}_i \mid \mathcal{D}^m)\right]\right).$$

Combining all of them together, we can write the bias of the predicted CATE as

$$\begin{aligned}
&\mathbb{E}\left[\widehat{\tau}(\widetilde{\mathbf{x}}_{\text{new}} \mid \widetilde{\mathcal{D}}) - \widehat{\tau}(\mathbf{x}_{\text{new}} \mid \mathcal{D})\right] = \\
&= \sum_{i \in \mathcal{I}^o:\ W_i = 1} \mathbb{E}[A_i - B_i - C_i] - \sum_{i \in \mathcal{I}^o:\ W_i = 0} \mathbb{E}[A_i - B_i - C_i] = \sum_{i \in \mathcal{I}^o} (2W_i - 1)\mathbb{E}[A_i - B_i - C_i] \\
&= \sum_{i \in \mathcal{I}^o} (2W_i - 1) \left\{ \mathbb{E}[Y_i - \widehat{m}^w(\mathbf{X}_i \mid \mathcal{D}^m)]\, \Delta^\ell_i(\mathbf{x}_{\text{new}}) - \mathbb{E}\left[\widehat{\ell}^w_i(\mathbf{x}_{\text{new}}, \mathcal{D}^\ell)\right] \cdot \Delta^m - \Delta^\ell_i(\mathbf{x}_{\text{new}}) \cdot \Delta^m_i \right\},
\end{aligned}$$

where $\Delta^\ell_i(\mathbf{x}_{\text{new}}) = \sum_{k=1}^{K} \frac{1}{k!} \sigma_k \text{trace}\left(\mathbb{E}\left[\partial^k_{\mathbf{x}_{\text{new}}, \mathcal{D}^\ell} \widehat{\ell}^{W_i}_i(\mathbf{x}_{\text{new}} \mid \mathcal{D}^\ell)\right]\right)$ represents the expected impact of LDP noise on the weight function, and $\Delta^m_i = \sum_{k=1}^{K} \frac{1}{k!} \sigma_k \text{trace}\left(\mathbb{E}\left[\partial^k_{\mathbf{X}_i, \mathcal{D}^m} \widehat{m}^w(\mathbf{X}_i \mid \mathcal{D}^m)\right]\right)$ represents the expected impact of LDP noise on the adjustment function. ■

Theorem B.1 formally demonstrates the impact of LDP protection on the bias properties of common CATE estimators. Specifically, LDP noise influences how the CATE model weights each individual i in the experiment set when forming the CATE prediction, $\Delta^\ell_i(\mathbf{x}_{\text{new}})$. This effect is

determined by two primary factors: (i) the scale of noise introduced into the data, denoted as σ_k, and (ii) the expected sensitivity of the weight functions to noise in $\Delta_i^{\ell}(\mathbf{x}_{\text{new}})$. Notably, when a model exhibits high complexity and nonlinearity, its expected sensitivity to noise increases, as higher-order derivatives can be large. This suggests that sophisticated machine learning models may experience significant bias when covariates are protected by LDP, particularly due to their inherent nonlinearity. Additionally, the impact on weights is heterogeneous across different covariate values, as $\Delta_i^{\ell}(\mathbf{x}_{\text{new}})$ is a function of $\mathbf{x}_{\text{new}}$. On the other hand, LDP noise also affects the adjustment function in the same manner (Δ_i^{m}), which can further amplify the bias.

Note that this theorem can be easily extended to cases where noise distributions differ by adjusting the expected sensitivity of the weight and adjustment functions. For example, the expected sensitivity of the weight function can be rewritten as:

$$\mathbb{E}\left[\widehat{\ell}_i^{w}(\widetilde{\mathbf{x}}_{\text{new}} \mid \widetilde{\mathcal{D}}^{\ell}) - \widehat{\ell}_i^{w}(\mathbf{x}_{\text{new}} \mid \mathcal{D}^{\ell})\right] = \sum_{k=1}^{K} \frac{1}{k!} \text{trace}\left(\mathbb{E}\left[\partial^{k}_{\mathbf{x}_{\text{new}},\mathcal{D}^{\ell}} \widehat{\ell}_i^{w}(\mathbf{x}_{\text{new}} \mid \mathcal{D}^{\ell})\right] \mathbb{E}\left[\text{diag}(\boldsymbol{\eta}^{\circ k}_{\mathbf{x}_{\text{new}},\mathcal{D}^{\ell}})\right]\right),$$

Then, substituting the expected sensitivity of the weight and adjustment functions into the proof above yields the bias for cases where the injected noise for each variable follows a different distribution.

B.2.2 Variance Analysis

Next, we investigate the effect of LDP covariate noise injection on the variance of predicted CATEs. As with our prior analysis, our goal is to identify and characterize the factors that influence this variance. This leads us to the following theorem:

Theorem B.2 ***(Variance Analysis)***

Under the same setting and assumptions in Theorem B.1, the impact of LDP protection on variance of the

predicted CATE of a given $\mathbf{x}_{new}$ can be written as:

$$
\begin{aligned}
&Var\left[\widehat{\tau}(\widetilde{\mathbf{x}}_{new})\right] - Var\left[\widehat{\tau}(\mathbf{x}_{new})\right] \\
&\underbrace{\sum_{k=1}^{K} \frac{1}{(k!)^2}\sigma_k^2 Var\left[\Delta_{\widehat{\tau}}^k(\mathbf{x}_{new})\right] + \sum_{k=1}^{K} \frac{1}{k!} Var\left[\eta_{i,p}^k\right] \mathbb{E}\left[\Gamma_{\widehat{\tau}}^k(\mathbf{x}_{new})\right]}_{(A)} + \\
&\underbrace{2\sum_{k=1}^{K} \frac{1}{k!}\sigma_k \nu_k(\mathbf{x}_{new})}_{(B)} + \underbrace{\sum_{k_1 \neq k_2} \frac{1}{k_1!}\frac{1}{k_2!}\sigma_{k_1}\sigma_{k_2}\zeta_{k_1,k_2}(\mathbf{x}_{new})}_{(C)}.
\end{aligned}
\tag{B.2}
$$

Here, $\Delta_{\widehat{\tau}}^k(\mathbf{x}_{new}) = trace\left(\partial^k \widehat{\tau}(\mathbf{x}_{new})\right)$ captures the k-th order influence of covariate noise on the predicted CATE, and $\Gamma_{\widehat{\tau}}^k(\mathbf{x}_{new}) = trace\left(\partial^k \widehat{\tau}(\mathbf{x}_{new})^{\circ 2}\right)$ represents the squared perturbation of the predicted CATE caused by LDP noise.[2] *The term $\nu_k(\mathbf{x}_{new}) = Cov\left[\widehat{\tau}(\mathbf{x}_{new}), \Delta_{\widehat{\tau}}^k\right]$ captures the covariance between the predicted CATE value (estimated on non-private data) and the k-th order change of predicted CATEs due to injected noises. Lastly, $\zeta_{k_1,k_2}(\mathbf{x}_{new}) = Cov\left[trace\left(\partial^{k_1}\widehat{\tau}(\mathbf{x}_{new})\right), trace\left(\partial^{k_2}\widehat{\tau}(\mathbf{x}_{new})\right)\right]$ measures how much the different orders of predicted CATE differences are correlated with each other.*

Proof. By Law of Total Variance, we have the following variance decomposition:

$$
\mathrm{Var}[\widehat{\tau}(\widetilde{\mathbf{x}}_{\text{new}} \mid \widetilde{\mathcal{D}})] = \mathbb{E}_{\mathcal{D}}\left\{\mathrm{Var}_{\boldsymbol{\eta}}[\widehat{\tau}(\widetilde{\mathbf{x}}_{\text{new}} \mid \widetilde{\mathcal{D}}) \mid \mathcal{D}]\right\} + \mathrm{Var}_{\mathcal{D}}\left\{\mathbb{E}_{\boldsymbol{\eta}}[\widehat{\tau}(\widetilde{\mathbf{x}}_{\text{new}} \mid \widetilde{\mathcal{D}}) \mid \mathcal{D}]\right\}, \tag{B.3}
$$

where $\boldsymbol{\eta}$ denotes the noises injected into the data.

First, note that the Taylor approximation of $\widehat{\tau}(\widetilde{\mathbf{x}}_{\text{new}} \mid \widetilde{\mathcal{D}})$ is

$$
\widehat{\tau}(\widetilde{\mathbf{x}}_{\text{new}} \mid \widetilde{\mathcal{D}}) = \widehat{\tau}(\mathbf{x}_{\text{new}} \mid \mathcal{D}) + \sum_{k=1}^{K} \frac{1}{k!}\left[\partial^k_{\mathbf{x}_{\text{new}},\mathcal{D}}\widehat{\tau}(\mathbf{x}_{\text{new}} \mid \mathcal{D})\right]\boldsymbol{\eta}^{\circ k},
$$

For simplicity, we use ∂^k in the rest of the proof to denote $\partial^k_{\mathbf{x}_{\text{new}},\mathcal{D}}$. Then, we have

$$
\mathrm{Var}_{\eta}[\widehat{\tau}(\widetilde{\mathbf{x}}_{\text{new}} \mid \widetilde{\mathcal{D}}) \mid \mathcal{D}] = \mathrm{Var}_{\eta}\left[\widehat{\tau}(\mathbf{x}_{\text{new}} \mid \mathcal{D}) + \sum_{k=1}^{K} \frac{1}{k!}\left[\partial^k \widehat{\tau}(\mathbf{x}_{\text{new}} \mid \mathcal{D})\right]\boldsymbol{\eta}^{\circ k}\right] = \sum_{k=1}^{K} \frac{1}{k!}\mathrm{Var}\left[\eta_{i,p}^k\right] \mathrm{trace}\left(\partial^k \widehat{\tau}(\mathbf{x}_{\text{new}})^{\circ 2}\right)
$$

since $\widehat{\tau}(\mathbf{x}_{\text{new}} \mid \mathcal{D})$ is a constant given $\mathcal{D}$ and fixed $\mathbf{x}_{\text{new}}$. Therefore, the first term in B.3 can be

[2] $\circ 2$ denotes the Hadamard's (i.e., element-wise) square operator.

written as

$$\mathbb{E}_{\mathcal{D}}\left\{\text{Var}_{\boldsymbol{\eta}}[\widehat{\tau}(\widetilde{\mathbf{x}}_{\text{new}} \mid \widetilde{\mathcal{D}}) \mid \mathcal{D}]\right\} = \sum_{k=1}^{K} \frac{1}{k!}\text{Var}\left[\eta_{i,p}^{k}\right] \mathbb{E}\left[\text{trace}\left(\partial^{k}\widehat{\tau}(\mathbf{x}_{\text{new}})^{\circ 2}\right)\right].$$

Similarly, we can write the second term as

$$\text{Var}_{\mathcal{D}}\left\{\mathbb{E}_{\boldsymbol{\eta}}[\widehat{\tau}(\widetilde{\mathbf{x}}_{\text{new}} \mid \mathcal{D}) \mid \mathcal{D}]\right\} - \text{Var}_{\mathcal{D}}\left[\widehat{\tau}(\mathbf{x}_{\text{new}} \mid \mathcal{D})\right] = \text{Var}_{\mathcal{D}}\left\{\sum_{k=1}^{K} \frac{1}{k!}\sigma_k\left[\partial^{k}\widehat{\tau}(\mathbf{x}_{\text{new}} \mid \mathcal{D})\right]\right)\right\}$$

$$= \text{Var}_{\mathcal{D}}\left[\sum_{k=1}^{K} \frac{1}{k!}\sigma_k \text{trace}\left(\partial^{k}\widehat{\tau}(\mathbf{x}_{\text{new}} \mid \mathcal{D})\right)\right] + 2\sum_{k=1}^{K} \frac{1}{k!}\sigma_k \text{Cov}_{\mathcal{D}}\left\{\widehat{\tau}(\mathbf{x}_{\text{new}} \mid \mathcal{D}), \text{trace}\left(\partial^{k}\widehat{\tau}(\mathbf{x}_{\text{new}} \mid \mathcal{D})\right)\right\}$$

$$= \text{Var}_{\mathcal{D}}\left[\widehat{\tau}(\mathbf{x}_{\text{new}} \mid \mathcal{D})\right] + \sum_{k=1}^{K} \frac{1}{(k!)^2}\sigma_k^2 \text{Var}\left[\text{trace}\left(\partial^{k}\widehat{\tau}(\mathbf{x}_{\text{new}} \mid \mathcal{D})\right)\right] + \sum_{k_1 \neq k_2} \frac{1}{k_1!}\frac{1}{k_2!}\sigma_{k_1}\sigma_{k_2}A_{k_1,k_2} + 2\sum_{k=1}^{K} \frac{1}{k!}\sigma_k B_k,$$

where $A_{k_1,k_2} = \text{Cov}_{\mathcal{D}}\left[\text{trace}\left(\partial^{k_1}\widehat{\tau}(\mathbf{x}_{\text{new}} \mid \mathcal{D})\right), \text{trace}\left(\partial^{k_2}\widehat{\tau}(\mathbf{x}_{\text{new}} \mid \mathcal{D})\right)\right]$ denotes the comovement between k_1-th and $k-2$-th derivatives, and $B_k = \text{Cov}_{\mathcal{D}}\left\{\widehat{\tau}(\mathbf{x}_{\text{new}} \mid \mathcal{D}), \text{trace}\left(\partial^{k}\widehat{\tau}(\mathbf{x}_{\text{new}} \mid \mathcal{D})\right)\right\}$ denotes the comovement between the initial CATE prediction and its higher-order derivative.

Combining these two terms together, we have

$$\text{Var}\left[\widehat{\tau}(\widetilde{\mathbf{x}}_{\text{new}} \mid \widetilde{\mathcal{D}})\right] - \text{Var}_{\mathcal{D}}\left[\widehat{\tau}(\mathbf{x}_{\text{new}} \mid \mathcal{D})\right] =$$

$$\sum_{k=1}^{K} \frac{1}{k!}\text{Var}\left[\eta_{i,p}^{k}\right] \mathbb{E}\left[\text{trace}\left(\partial^{k}\widehat{\tau}(\mathbf{x}_{\text{new}})^{\circ 2}\right)\right] + \sum_{k=1}^{K} \frac{1}{(k!)^2}\sigma_k^2 \text{Var}\left[\text{trace}\left(\partial^{k}\widehat{\tau}(\mathbf{x}_{\text{new}} \mid \mathcal{D})\right)\right] +$$

$$\sum_{k_1 \neq k_2} \frac{1}{k_1!}\frac{1}{k_2!}\sigma_{k_1}\sigma_{k_2}A_{k_1,k_2} + 2\sum_{k=1}^{K} \frac{1}{k!}\sigma_k B_k.$$

■

Theorem B.2 demonstrates the mixed effects of LDP noise on the variance of the predicted CATE. In Equation (B.2), term (A) captures the increase in variance due to the additional randomness introduced by LDP protection. Since all the elements in (A) are strictly positive, this term is increasing with the magnitude of injected noise, specifically with respect to σ_k and $\text{Var}\left[\eta_{i,p}^{k}\right]$. Term (B) reflects the relationship between the predicted CATE estimated using non-private data (had it been available) and the influence of the injected noise on these predictions. If the injected noise tends to increase the predictions for individuals already showing a positive CATE (such that the covariance in (B) is positive), estimating CATE on injected noise will make even more extreme predictions. Consequently, LPD will further increase the variance of the

predicted CATE. On the other hand, if this covariance is negative, the added noise would draw the predictions towards zero, reducing the variance of the predicted CATE. Term (C) denotes the co-movement across k_1-th and k_2-th order shifts in predicted CATEs. If these shifts tend to move in the same direction (i.e., having positive covariance), the variance is increased with larger σ_k. Conversely, if the k_1-th and k_2-th order fluctuations are in opposite directions (i.e., having negative covariance), they will offset each other, leading to no increase in variance.

B.3 Proof for Proposition 2.1

Proof. *Proof*: First, we can rearrange the AIPW score as follows:

$$\check{\tau}_j = \underbrace{\left[\frac{W_j\widetilde{Y}_j}{e_j} - \frac{(1-W_j)\widetilde{Y}_j}{1-e_j}\right]}_{\text{IPW term}} + \underbrace{\left[\widehat{\mu}_1(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}) \cdot \left(1 - \frac{W_j}{e_j}\right) - \widehat{\mu}_0(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}) \left(1 - \frac{1-W_j}{1-e_j}\right)\right]}_{\text{Noise term}},$$

where $\widehat{\mu}_1(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}})$ and $\widehat{\mu}_0(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}})$ are the estimated conditional mean outcome models for the treatment and control individuals.

For the IPW term, we have its conditional expectation being

$$\begin{aligned}
&\mathbb{E}\left[\frac{W_j\widetilde{Y}_j}{e_j} - \frac{(1-W_j)\widetilde{Y}_j}{1-e_j} \mid \mathbf{X}_j\right] \\
&\quad= \underbrace{\mathbb{P}(W_j = 1 \mid \mathbf{X}_j)}_{=e_j} \cdot \mathbb{E}\left[\frac{\widetilde{Y}_j}{e_j} \mid W_j = 1, \mathbf{X}_j\right] + \underbrace{\mathbb{P}(W_j = 0 \mid \mathbf{X}_j)}_{=1-e_j} \cdot \mathbb{E}\left[\frac{\widetilde{Y}_j}{1-e_j} \mid W_j = 0, \mathbf{X}_j\right] \\
&\quad= \mathbb{E}[\widetilde{Y}_j \mid W_j = 1, \mathbf{X}_j] - \mathbb{E}[\widetilde{Y}_j \mid W_j = 0, \mathbf{X}_j] \\
&\quad= \mathbb{E}[Y_j \mid W_j = 1, \mathbf{X}_j] - \mathbb{E}[Y_j \mid W_j = 0, \mathbf{X}_j] \quad \text{(since the additive noise has mean zero)} \\
&\quad= \mathbb{E}[Y_j(1) \mid \mathbf{X}_j] - \mathbb{E}\left[Y_j(0) \mid \mathbf{X}_j\right] = \tau(\mathbf{X}_j) \quad \text{(by the unconfoundedness assumption).}
\end{aligned}$$

For the noise term, we have its conditional expectation being

$$\begin{aligned}
&\mathbb{E}\left[\widehat{\mu}_1(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}) \cdot \left(1 - \tfrac{W_j}{e_j}\right) - \widehat{\mu}_0(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}) \left(1 - \tfrac{1-W_j}{1-e_j}\right) \mid \mathbf{X}_j\right] = \\
&\qquad \mathbb{E}\left[\widehat{\mu}_1(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}) \mid \mathbf{X}_j\right] \cdot \underbrace{\left(1 - \tfrac{\mathbb{E}[W_j|\mathbf{X}_j]}{e_j}\right)}_{=0 \text{ since } \mathbb{E}[W_j|\mathbf{X}_j]=e_j} + \mathbb{E}\left[\widehat{\mu}_0(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}) \mid \mathbf{X}_j\right] \cdot \underbrace{\left(1 - \tfrac{1-\mathbb{E}[W_j|\mathbf{X}_j]}{1-e_j}\right)}_{=0 \text{ since } \mathbb{E}[W_j|\mathbf{X}_j]=e_j} = 0.
\end{aligned}$$

Combining these two results, we have $\mathbb{E}\left[\check{\tau}_j \mid \mathbf{X}_j\right] = \tau(\mathbf{X}_j)$. Therefore, the expected approximation error is zero since

$$\mathbb{E}\left[\check{\epsilon}_j\right] = \mathbb{E}\left[\mathbb{E}\left[\check{\tau}_j - \tau(\mathbf{X}_j) \mid \mathbf{X}_j\right]\right] = 0.$$

■

B.4 Additional Details for the Simulation Analyses in Section 3

In this appendix, we provide model implementation details for the simulation analysis in Section 2.3. Specifically, we implement the following three CATE models:

1. *Causal Forest* [199]: We implement causal forest using the `grf` package with default parameters.

2. *R-learner with XGBoost models* [155]: We implement R-learner with XGBoost models for both the conditional mean outcome and CATE models. Hyperparameter tuning follows the methods implemented by the R-package `rlearner` provided by [155].

3. *T-learner with XGBoost models*: We implement the T-learner with correctly specified models. Specifically, for each treatment condition, we estimate a regression model of the outcome on covariates using second-order polynomial specifications. The CATE is then calculated by subtracting the predicted outcome from the control group model from that of the treatment group model.

B.5 Additional Details and Results for the Simulation Analyses in Section 5

In this appendix, we provide implementation details about the simulation analyses described in Section 2.5 of the main document and present robustness checks.

B.5.1 Details on Model Specifications

Constructing the Initial CATE Model

In our main analysis, we implement the R-learner [155] using regression forests from the `grf` package with default parameters for both the conditional mean outcome models and the CATE model. The true propensity score is used to construct Robinson's transformation. In Electronic Companion B.5.6, we provide additional robustness checks for different initial CATE models.

Constructing the AIPW Score for Post-Processing

To construct the AIPW score, we use a fixed propensity score ($e = 0.5$) since the treatment assignment is completely randomized. We evaluate three candidate models for the conditional mean: linear regression, linear regression with interactions, and regression forests. The training set consists of 1,000 individuals (500 treated and 500 non-treated), while the holdout evaluation set includes 2,000 individuals (1,000 treated and 1,000 non-treated). Table B.3 reports the mean squared error (MSE) of each model in predicting the outcome. Since linear regression with interactions yields the lowest prediction error, we select it as our preferred model for the AIPW score. Additionally, we use the same model to construct calibrators during the calibration process. Notably, the conditional mean model used in the AIPW transformation and calibration process does not align exactly with the data-generating process. However, the algorithm still effectively reduces the MSE of the CATE predictions.

Table B.3: *MSE of Conditional Outcome Models Across Varying Privacy Levels*

(a) Scenario 1: LDP-Protected Covariates

σ	Linear Regression	Regression with Interactions	Regression Forest
0.0	33.4	29.6	33.1
0.2	34	30.6	34.2
0.4	36.1	33.7	36.3
0.6	39.2	37.8	39.7
0.8	41.4	40.9	41.3
1.0	43.4	43.0	43.1

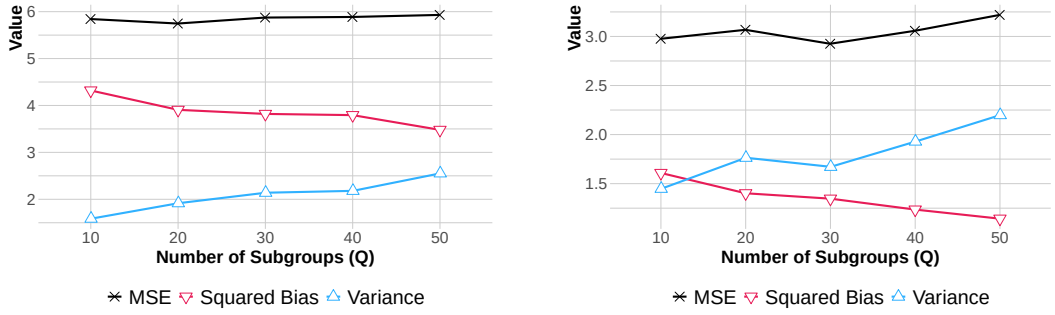
(b) Scenario 2: LDP-Protected Outcome

σ	Linear Regression	Regression with Interactions	Regression Forest
0	33.4	29.5	33.6
1	35.1	31.1	35
2	41.3	37.5	41.4
3	51.6	48.6	51.7
4	65.8	63.3	65.6
5	83.9	81.7	84.4

Determine Number of Subgroups and Number of Iterations

Next, we investigate the sensitivity of our solution to the pre-specified number of subgroups (Q) and the number of iterations (R). We start by illustrating the bias-variance trade-offs for the number of subgroups. Figure B.1 displays the predictive error metrics when applying the proposed solution with varying values of Q under two scenarios with high privacy protection. As we increase the number of subgroups, the bias decreases for both scenarios, while the variance correspondingly increases. This outcome showcases a classic bias-variance trade-off commonly

observed in machine learning models: increasing model complexity may enhance precision but concurrently introduce additional variability. Notably, when the outcome is protected by LDP, changes in Q do not significantly affect overall accuracy. On the other hand, when the covariates are protected by LDP, using the smallest value of Q achieves the lowest MSE.

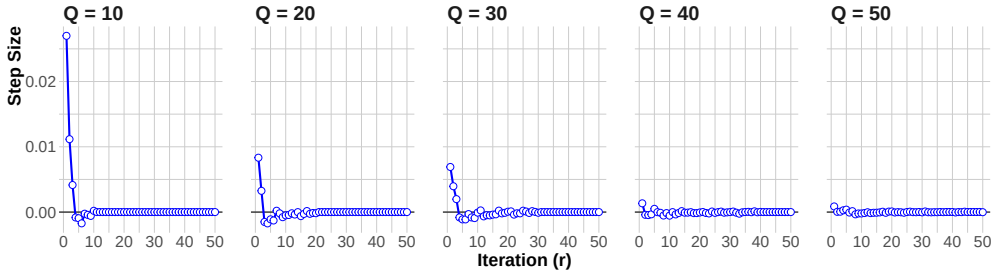


(a) *Scenario 1: LDP-Protected Covariates*

(b) *Scenario 2: LDP-Protected Outcome*

Figure B.1: *Predictive Errors of Proposed Method with Varying Numbers of Subgroups*

Figure B.2 illustrates the step size $\rho_{q^\star}^{[r]}$ in each iteration (capped at 50 iterations) when applying the proposed solution with varying numbers of subgroups (i.e., varying values of Q). Generally, the step sizes approach zero before the completion of the first Q iterations, regardless of the chosen value for Q. This observation suggests that performing $R = Q$ iterations is typically sufficient for the proposed solution, thereby providing a practical guideline for setting this hyperparameter.



Note: Each point is calculated by averaging the mean step size over 100 bootstrap replications. We present results from the case when the covariates are protected by LDP with $\sigma = 1.0$, but the pattern remains the same across different scenarios. We use R-learner with regression forests as the initial CATE model.

Figure B.2: *Step Size in Each Iteration*

B.5.2 Additional Performance Evaluation

Bias and Variance.

To further examine how the PROPOSED method enhances predictive accuracy, we extend the analysis in Section 2.5 by decomposing the mean squared error (MSE) into its squared bias and variance components. These metrics are computed as follows:

1. *(Mean Squared Bias)* We calculate the average deviation of model predictions from the actual CATEs across individuals in the holdout set, i.e.,

$$\widehat{\mathbb{E}}^2\left[(\widehat{err}_i^b)\right] = \frac{1}{N_{\text{holdout}}} \sum_{i \in D_{\text{holdout}}} \left[\frac{1}{B}\sum_{b=1}^{B} \widehat{err}_i^b\right]^2.$$

2. *(Mean Variance)* We calculate the average variability in the model's predictions across individuals in the holdout set when the model is trained with different data sets, i.e.,

$$\widehat{\text{Var}}\left[\widehat{\tau}\right] = \frac{1}{B}\sum_{b=1}^{B} \frac{1}{N_{\text{holdout}}} \sum_{i \in D_{\text{holdout}}} \left\{\left(\widehat{\tau}^b(\mathbf{X}_i) - \widehat{\mathbb{E}}\left[\widehat{\tau}^b(\mathbf{X}_l)\right]\right)^2\right\}.$$

Figure B.3 presents the mean squared error, squared bias, and variance of different methods across varying sample sizes. First, we find that the NON-HONEST method achieves the smallest squared bias because its initial CATE model is trained on a sample three times larger than that used by the PROPOSED and SPLIT-ONLY methods. However, as the sample size increases, the squared bias of the PROPOSED method decreases to a level comparable to that of the NON-HONEST method, whereas the Honest method does not exhibit the same improvement. Second, despite its lower bias, the NON-HONEST method consistently shows significantly higher variance than the other methods. This is due to overfitting to noise in the AIPW score, as discussed in Section 2.4.3. Together, these findings confirm the importance of the honesty principle in mitigating overfitting when the AIPW score is noisy.

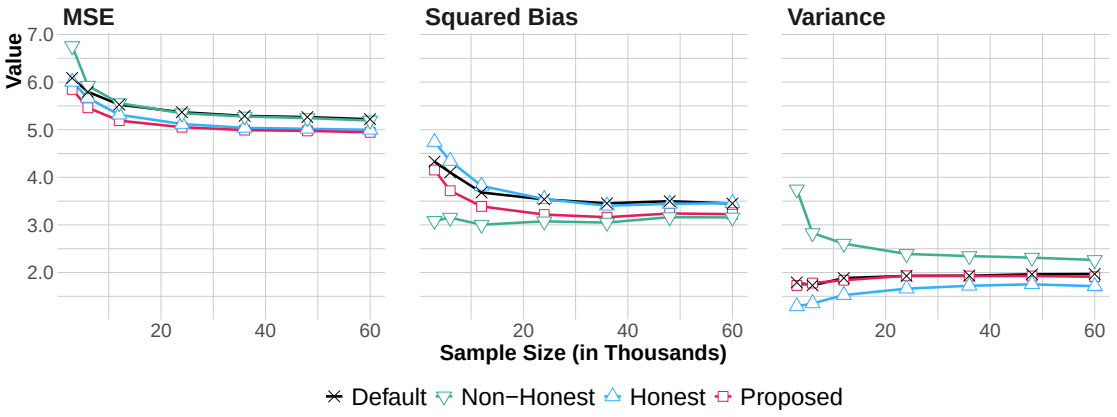
Value Improvement.

We also assess the improvement in targeting performance achieved by our proposed method when decision-makers rely on CATE models estimated from LDP-protected data. Specifically, we

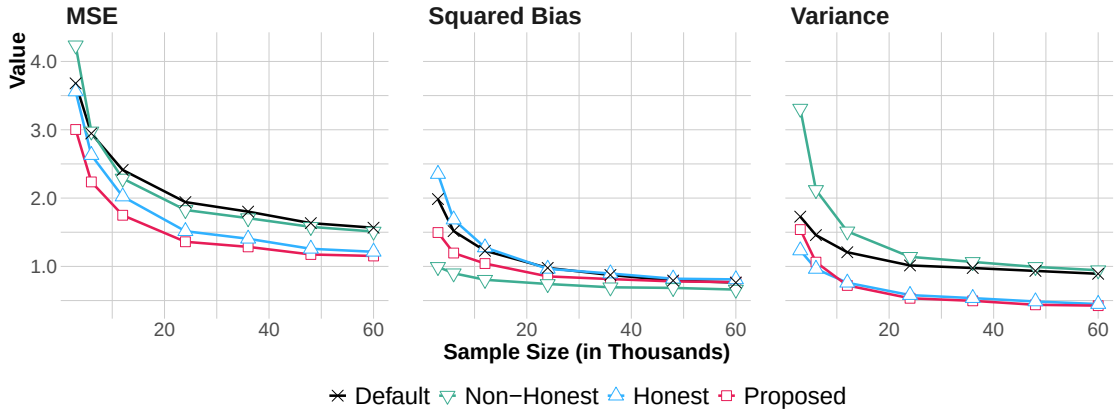
measure the incremental value gained from targeting individuals with positive predicted CATEs:

$$V(\widehat{\tau}) = \frac{1}{B}\sum_{b=1}^{B}\left[\frac{1}{N_{\text{holdout}}}\sum_{l\in\widetilde{\mathcal{D}}_{\text{holdout}}}\tau(\mathbf{X}_l)\cdot\mathbf{1}(\widehat{\tau}_l^b > 0)\right],$$

where $\widehat{\tau}_l^b$ is the predicted CATE of individual l in the b-th bootstrap replication. We then report the value improvement of a method $\widehat{\tau}$ compared to the DEFAULT method $\widehat{\tau}^{\text{DEFAULT}}$ as the percentage difference in their targeting value, i.e., $100\% \times \frac{V(\widehat{\tau})-V(\widehat{\tau}^{\text{DEFAULT}})}{V(\widehat{\tau}^{\text{DEFAULT}})}$.



(a) ***Scenario 1: LDP-Protected Covariates***



(b) ***Scenario 2: LDP-Protected Outcome***

Note: We simulate 100 replications to compute the bootstrap mean MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point. The results presented here are based on the use of R-learner R-learner with regression forests as the initial CATE model.

Figure B.3: *Squared Bias and Variance with Varying Sample Sizes*

Varying Privacy Levels. Table B.4 presents the value improvement of different methods over the Default method. Specifically, we report both the percentage improvement and the proportion of bootstrap replications in which the focal method outperforms the Default approach. The Proposed achieves significant positive value improvement regardless of privacy levels. Notably, the relative value improvement of the Proposed methods over the Default is even greater at higher privacy levels. In contrast, both the Split-only and Non-Honest methods did not improve targeting value significantly, and the Non-Honest approach even results in value loss when covariates are protected by high levels of noise. Together, these findings suggest that the Subgroup Cross-Learning approach is particularly effective in addressing the challenges of stricter privacy protection settings, providing robust value improvements in targeting.

Table B.4: *Value Improvement Across Varying Privacy Levels*

(a) Scenario 1: LDP-Protected Covariates

σ	Proposed	Split-only	Non-Honest
0	7.0% (98%)	2.4% (64%)	2.8% (76%)
0.2	8.4% (98%)	2.5% (74%)	2.5% (64%)
0.4	10.6% (100%)	3.1% (70%)	2.8% (66%)
0.6	13.6% (100%)	0.6% (52%)	1.7% (50%)
0.8	18.6% (98%)	1.6% (52%)	1.5% (52%)
1.0	14.9% (84%)	2.6% (54%)	-10.3% (32%)

(b) Scenario 2: LDP-Protected Outcome

σ	Proposed	Split-only	Non-Honest
0	7.0% (98%)	2.4% (64%)	2.8% (76%)
1	7.9% (98%)	1.2% (67%)	3.2% (69%)
2	8.8% (99%)	1.1% (66%)	3.5% (79%)
3	10.8% (96%)	0.9% (53%)	3.6% (66%)
4	11.5% (94%)	1.9% (62%)	1.7% (58%)
5	13.7% (94%)	2.6% (65%)	1.8% (54%)

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses). The results presented here are based on the use of R-learner as the initial CATE model. Results derived from different CATE models are available in Electronic Companion B.5.6.

Varying Sample Sizes. Table B.5 reports the value improvement across different sample sizes. Consistent with the findings in Section 2.5.5, the Proposed method consistently outperforms other approaches, achieving the best performance across all sample sizes. When the sample size is large, the Split-only method performs similar the Proposed method.

Runtime Analysis

A potential concern with the post-processing approach is the additional time required for implementation. Table B.6 summarizes the runtime of each method discussed in Section 2.5.2 across different sample sizes. Overall, we find that while post-processing methods take longer

Table B.5: *Value Improvement for Different Methods Across Varying Sample Sizes*

(a) Scenario 1: LDP-Protected Covariates

Sample Size	Proposed	Split-only	Non-honest
3,000	14.9% (84%)	2.6% (54%)	−10.3% (32%)
6,000	17.5% (96%)	3.4% (68%)	0.7% (52%)
12,000	13.7% (100%)	8.0% (92%)	0.1% (50%)
24,000	11.5% (100%)	9.9% (100%)	1.1% (63%)
36,000	10.8% (100%)	9.1% (100%)	0.9% (65%)
48,000	10.6% (100%)	9.5% (100%)	1.3% (70%)
60,000	11.4% (100%)	8.8% (100%)	1.5% (68%)

(b) Scenario 2: LDP-Protected Outcome

Sample Size	Proposed	Split-only	Non-honest
3,000	13.8% (94%)	0.1% (52%)	1.0% (56%)
6,000	10.7% (100%)	5.6% (88%)	2.8% (74%)
12,000	8.3% (100%)	5.3% (94%)	2.6% (82%)
24,000	6.6% (100%)	5.1% (100%)	2.2% (100%)
36,000	5.6% (100%)	4.9% (100%)	1.6% (100%)
48,000	5.5% (100%)	4.5% (100%)	1.4% (88%)
60,000	4.2% (100%)	3.6% (100%)	1.2% (78%)

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses). The results presented here are based on the use of R-learner as the initial CATE model. Results derived from different CATE models are available in Electronic Companion B.5.6.

than the Default approach, the additional time remains manageable. For instance, with a sample size of 60,000 individuals, the Proposed method adds only 22 extra seconds compared to the Default method.

Table B.6: *Runtime (in Seconds) of Different Methods Across Varying Sample Sizes*

Sample Size	Proposed	Split-only	Non-Honest	Default
3k	8.3 (0.3)	7.0 (0.2)	6.9 (0.2)	5.9 (0.2)
6k	16.0 (0.8)	12.9 (0.7)	12.9 (0.7)	11.6 (0.6)
12k	30.7 (1.0)	25.7 (0.8)	25.6 (0.8)	24.6 (0.9)
24k	73 (1.3)	51.8 (0.8)	51.6 (0.5)	53.9 (1.6)
36k	112.5 (1.4)	80.3 (1.1)	80.2 (1.1)	92.8 (3.0)
48k	158.6 (8.2)	114.3 (3.9)	114.1 (3.9)	135.7 (6.9)
60k	197.7 (4.9)	144.4 (3.8)	144.3 (3.7)	175.3 (5.8)

Note: We calculate the mean running time from 100 simulation replications, along with the standard deviation in parentheses.

B.5.3 Policy Learning as An Alternative Benchmark

In this appendix, we explore whether the existing policy learning method [17] could serve as a potential solution for mitigating LDP noise, as learning a mapping from covariates to a binary decision may be easier than estimating the true CATEs. To implement this approach, we first generate the AIPW score using the procedure described earlier. We then derive the targeting policy based on this score by solving the optimization problem:

$$\widehat{\pi} = \arg\max_{\pi} \sum_{i \in \mathcal{D}_{\text{train}}} [2\pi(\mathbf{X}_i) - 1]\, \widehat{\Gamma}_i,$$

where we constrain π in the class of probability forest in the `grf` package.

Table B.7 compares the value improvement (as defined in Section 2.5.4) of policy learning

relative to targeting individuals with positive predicted CATEs from the DEFAULT method. Overall, we find that policy learning does not improve targeting performance compared to the DEFAULT method. We hypothesize two possible reasons for this result. First, the AIPW score used in policy learning is noisier than the predicted CATEs, making it difficult to learn an effective targeting policy. Second, when directly learning the targeting policy, covariate noise complicates the ability to establish an optimal decision rule.

Table B.7: *Value Improvement Across Varying Privacy Levels*

(a) Scenario 1: LDP-Protected Covariates

σ	Proposed	Policy Learning
0	7.0% (98%)	−1.6% (22%)
0.2	8.4% (98%)	−2.6% (9%)
0.4	10.6% (100%)	−1.3% (39%)
0.6	13.6% (100%)	−1.2% (35%)
0.8	18.6% (98%)	−2.0% (39%)
1	14.9% (84%)	0.2% (65%)

(b) Scenario 2: LDP-Protected Outcome

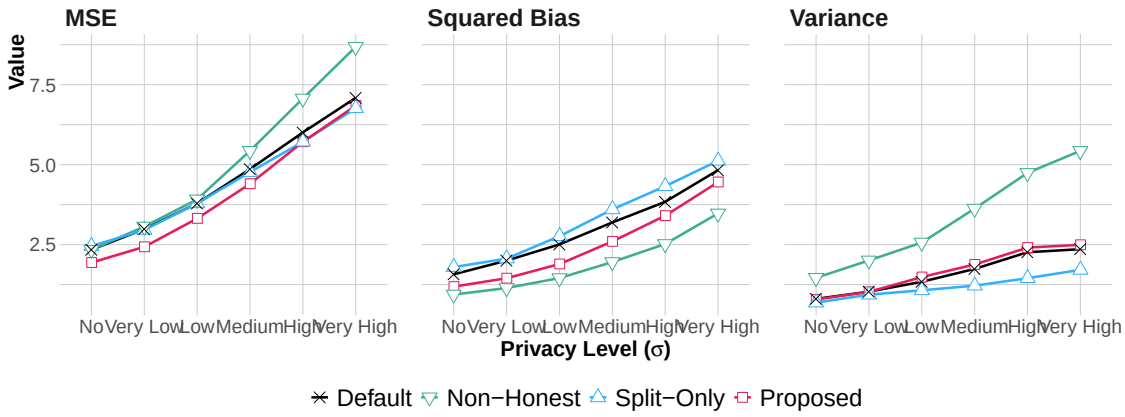
σ	Proposed	Policy Learning
0	7.0% (98%)	−1.6% (22%)
1	7.9% (98%)	−2.0% (26%)
2	8.8% (99%)	−1.6% (30%)
3	10.8% (96%)	−1.1% (31%)
4	11.5% (94%)	−1.4% (35%)
5	13.4% (94%)	−0.1% (43%)

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses). The results presented here are based on the use of R-learner with regression forests as the initial CATE model.

B.5.4 Results with LDP Protection Applied to Both Outcomes and Covariates

In this appendix, we present results where both the outcome and covariates are protected under Local Differential Privacy (LDP). Specifically, we consider five privacy levels—very low, low, medium, high, and very high—by varying the scale of Laplace noise: $\sigma \in \{1, 2, 3, 4, 5\}$ for outcomes and $\sigma \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$ for covariates. We then replicate the previous analyses under these settings.

Varying Privacy Levels. Figure B.4 presents the MSE, squared bias, and variance of different methods across different privacy levels, where we apply R-learners with regression forests as the DEFAULT method and the initial CATE models. The results are consistent with the findings from the main analysis. Table B.8 reports the AUTOC values across different privacy levels. The PROPOSED method consistently outperforms other approaches, achieving the best performance across all privacy levels. Table B.9 presents the value improvement of different methods across various privacy levels. The results align with the findings from the main analysis.



Note: We simulate 100 replications to compute the bootstrap mean MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point. The results presented here are based on the use of R-learner with regression forests as the initial CATE model.

Figure B.4: *Statistical Accuracy with Varying Privacy Levels*

Table B.8: *AUTOC Values for Different Methods Across Varying Privacy Levels*

Privacy	Proposed	Honest	Non-Honest	Default
No	2.37 (96%)	2.34 (75%)	2.33 (48%)	2.33
Very Low	2.3 (100%)	2.26 (76%)	2.22 (52%)	2.22
Low	2.14 (100%)	2.11 (83%)	2.08 (75%)	2.06
Medium	1.92 (100%)	1.9 (96%)	1.81 (46%)	1.81
High	1.64 (92%)	1.66 (97%)	1.54 (57%)	1.55
Very High	1.36 (83%)	1.40 (92%)	1.26 (48%)	1.28

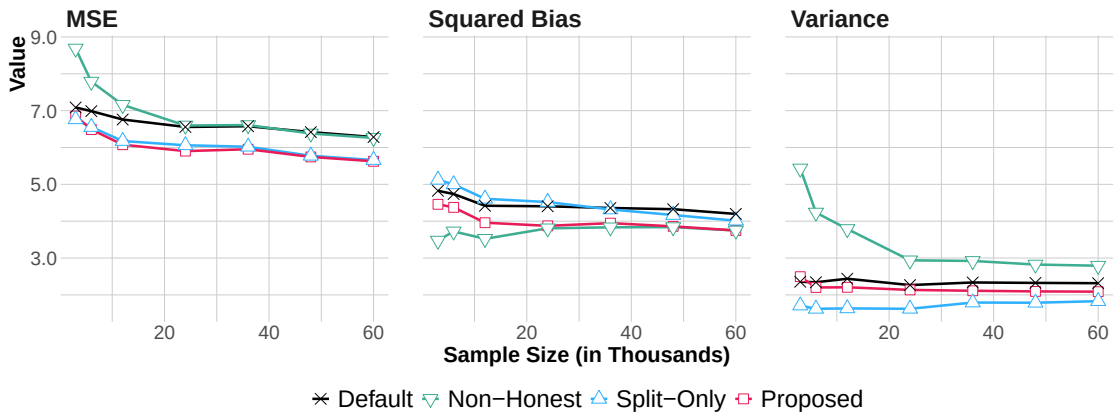
Note: We calculate the AUTOC value from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses). The results presented here are based on the use of R-learner with regression forests as the initial CATE model.

Table B.9: *Value Improvement Across Varying Privacy Levels*

Privacy	Proposed	Honest	Non-Honest
No	5.3% (84%)	−1.4% (36%)	1.0% (64%)
Very Low	8.8% (100%)	0.7% (56%)	0.2% (56%)
Low	12.5% (100%)	1.1% (56%)	4.5% (64%)
Medium	15.3% (100%)	1.1% (56%)	−4.3% (40%)
High	14.2% (84%)	6.4% (72%)	−18.1% (24%)
Very High	24.7% (72%)	21.8% (68%)	−68.2% (12%)

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses). The results presented here are based on the use of R-learner with regression forests as the initial CATE model.

Varying Sample Sizes. Figure B.5 presents the MSE, squared bias, and variance of different methods across different privacy levels, where we apply Casual Forest as the DEFAULT method and the initial CATE models. The results are consistent with the findings from the main analysis.



Note: We simulate 100 replications to compute the bootstrap mean MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point. The results presented here are based on the use of R-learner with regression forests as the initial CATE model.

Figure B.5: *Statistical Accuracy with Varying Sample Sizes*

Table B.10 reports the AUTOC values across different sample sizes. Table B.11 presents the value improvement of different methods across various sample sizes. The results align with the findings in Section 2.5.5.

Table B.10: *AUTOC Values for Different Methods Across Varying Sample Sizes*

Sample Size	Proposed	Honest	Non-Honest	Default
3k	1.36 (83%)	1.4 (92%)	1.26 (48%)	1.28
6k	1.45 (96%)	1.42 (92%)	1.33 (76%)	1.30
12k	1.57 (100%)	1.54 (100%)	1.45 (88%)	1.40
24k	1.6 (100%)	1.55 (100%)	1.47 (100%)	1.42
36k	1.61 (100%)	1.58 (100%)	1.49 (100%)	1.45
48k	1.62 (100%)	1.60 (100%)	1.50 (100%)	1.46
60k	1.65 (100%)	1.64 (100%)	1.55 (100%)	1.50

Note: We calculate the AUTOC value from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses). The results presented here are based on the use of R-learner with regression forests as the initial CATE model.

B.5.5 Additional Benchmarks for Post-processing

In this appendix, we provide additional benchmarks for post-processing to further demonstrate the value of our PROPOSED solution. Specifically, we evaluate (i) different subgroup partitioning

Table B.11: *Value Improvement for Different Methods Across Varying Sample Sizes*

Sample Size	Proposed	Honest	Non-Honest
3k	24.7% (72%)	21.8% (68%)	−68.2% (12%)
6k	75.8% (92%)	74.6% (92%)	−50.6% (20%)
12k	81.6% (100%)	77.9% (100%)	-13.5% (24%)
24k	67.5% (100%)	55.5% (100%)	3.8% (65%)
36k	86.1% (100%)	83.9% (100%)	1.7% (72%)
48k	70.2% (100%)	74% (100%)	11.2% (80%)
60k	57% (100%)	60.4% (100%)	4.4% (80%)

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses). The results presented here are based on the use of R-learner with regression forests as the initial CATE model.

strategies within the PROPOSED method to assess the impact of within-group homogeneity in the subgroup cross-learning strategy and (ii) stochastic gradient boosting [78] for honest post-processing to compare the effectiveness of our proposed subgroup cross-learning approach against a standard machine learning technique.

Alternative Subgroup Partitioning Strategies.

We compare three alternative subgroup partitioning strategies to splitting individuals based on their predicted CATEs from the initial CATE predictions.

1. *Partitioning based on covariate values*: We use K-means clustering on covariate values to define the subgroups.

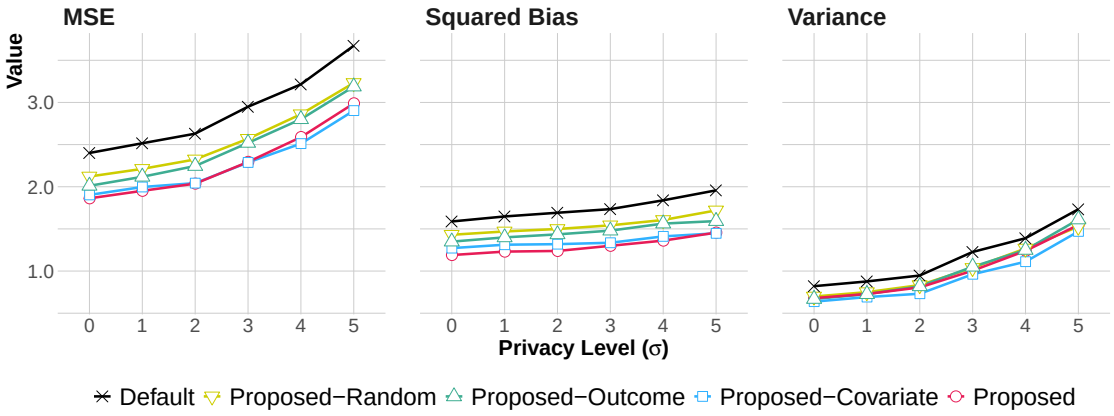
2. *Partitioning based on the outcome*: We categorize individuals based on their outcome variable.

3. *Partitioning randomly*: We generate random subgroups of individuals.

Figure B.6 presents the MSE, squared bias, and variance across different privacy levels with small samples ($N = 3,000$). Overall, all subgroup partitioning strategies improve accuracy compared to the DEFAULT method. Among them, subgrouping based on predicted CATEs and covariate values yields the best performance. Figure B.7 presents the MSE, squared bias, and variance across different sample sizes under high privacy levels. Overall, all subgroup partitioning strategies improve accuracy compared to the DEFAULT method. Additionally, when the sample size is large, performance remains similar across different subgroup partitioning strategies.

(a) *Scenario 1: LDP-Protected Covariates*

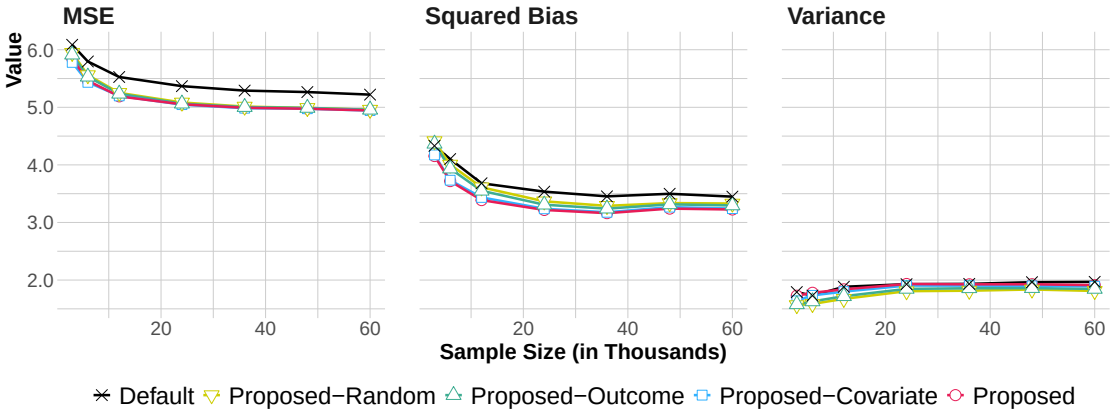
(b) *Scenario 2: LDP-Protected Outcome*



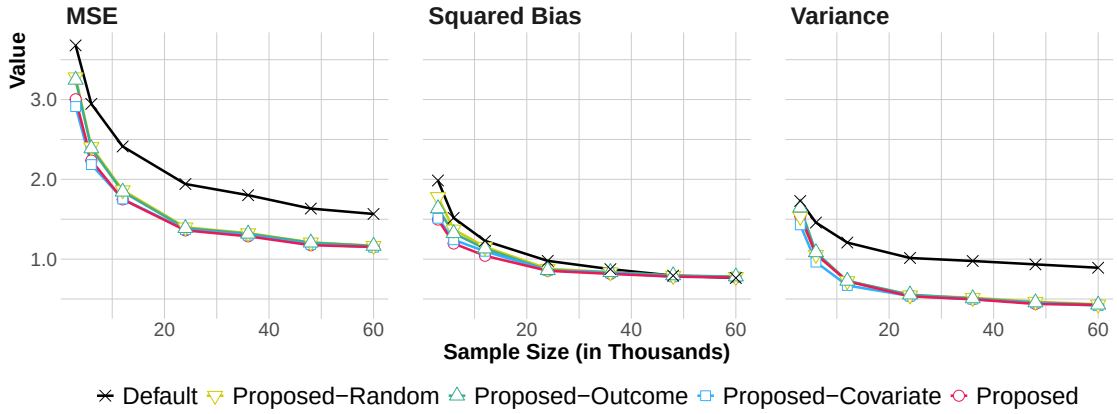
Note: We simulate 100 replications to compute the bootstrap mean MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point. The results presented here are based on the use of R-learner with regression forests as the initial CATE model.

Figure B.6: *Statistical Accuracy with Varying Privacy Levels Across Subgroup Partitioning Strategies*

(a) *Scenario 1: LDP-Protected Covariates*



(b) *Scenario 2: LDP-Protected Outcome*



Note: We simulate 100 replications to compute the bootstrap mean MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point. The results presented here are based on the use of R-learner with regression forests as the initial CATE model.

Figure B.7: *Statistical Accuracy with Varying Sample Sizes Across Subgroup Partitioning Strategies*

Alternative Post-processing Methods

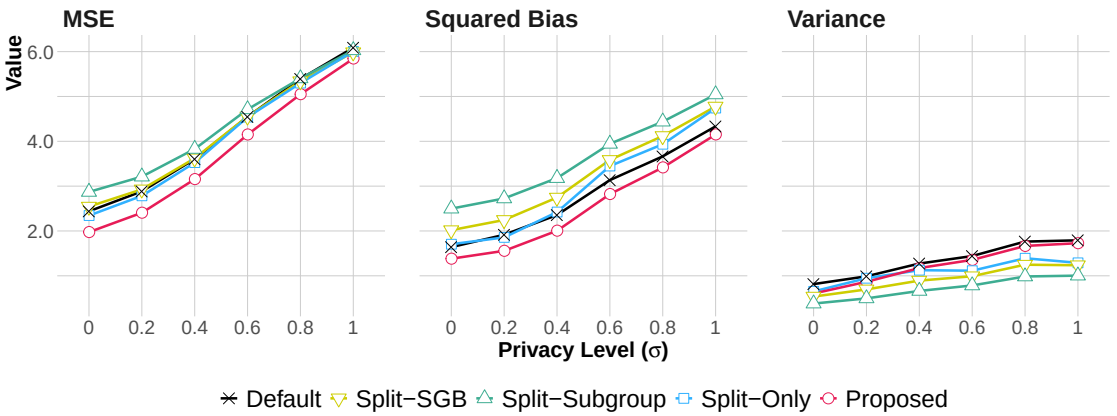
In this appendix, we compare several alternative boosting procedures for post-processing. We first consider the alternative honest post-processing methods.

1. *Post-Processing with Sample Splitting and Subgroup Updates* (Split-Subgroup): This method performs post-processing using a separate adjustment dataset $\widetilde{\mathcal{D}}_{\text{adj}}$ and stops when there is no improvement on a third validation set $\widetilde{\mathcal{D}}_{\text{val}}$. We perform subgroup learning, where subgroups are determined based on the predicted CATEs from the initial CATE model. However, the step size is determined within the *same* subgroup of individuals rather than using other subgroups.

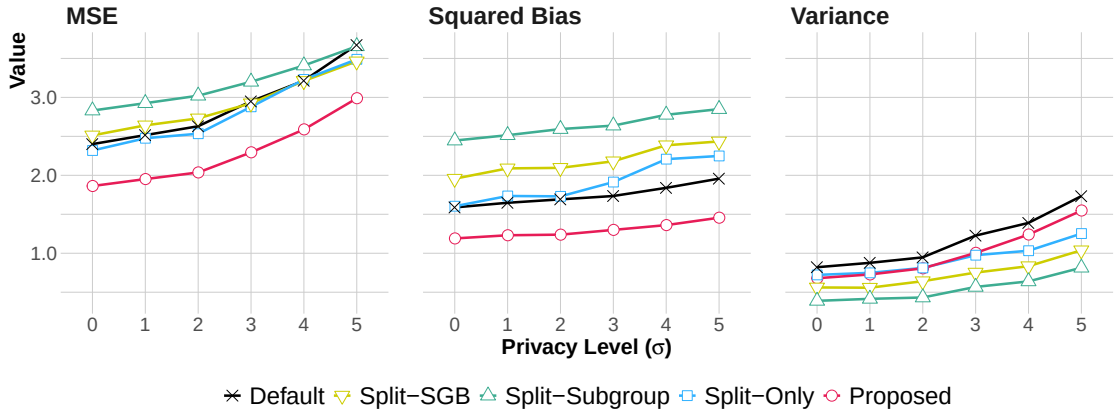
2. *Post-Processing with Sample Splitting and Subgroup Updates* (Split-SGB): This method performs post-processing using a separate adjustment dataset $\widetilde{\mathcal{D}}_{\text{adj}}$ and stops when there is no improvement on a third validation set $\widetilde{\mathcal{D}}_{\text{val}}$. In each iteration, we apply the stochastic gradient descent method to update the model, i.e., we randomly sample 20% of individuals from the calibration set, then use this random subsample to construct the adjustment model and determine the step size.

Figure B.8 presents the MSE, squared bias, and variance of different methods across different privacy levels. Overall, the alternative honest post-processing methods do not improve accuracy and instead exhibit significantly higher bias.

Next, we further examine the value of the proposed subgroup cross-learning in settings without sample splitting (Non-honest-Cross). Figure B.9 presents the MSE, squared bias, and variance of different methods across different privacy levels. Consistent with our theoretical insights, we find that subgroup cross-learning enhances predictive accuracy even without using a separate dataset for model post-processing.



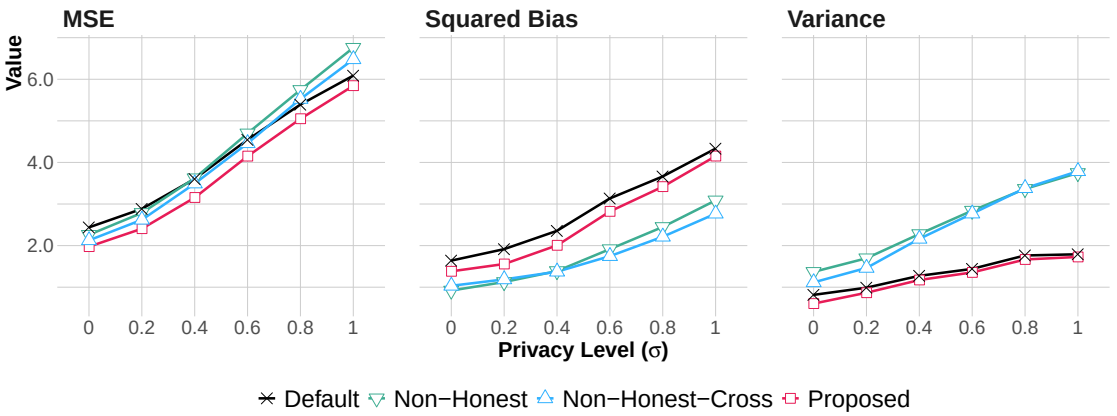
(a) ***Scenario 1: LDP-Protected Covariates***



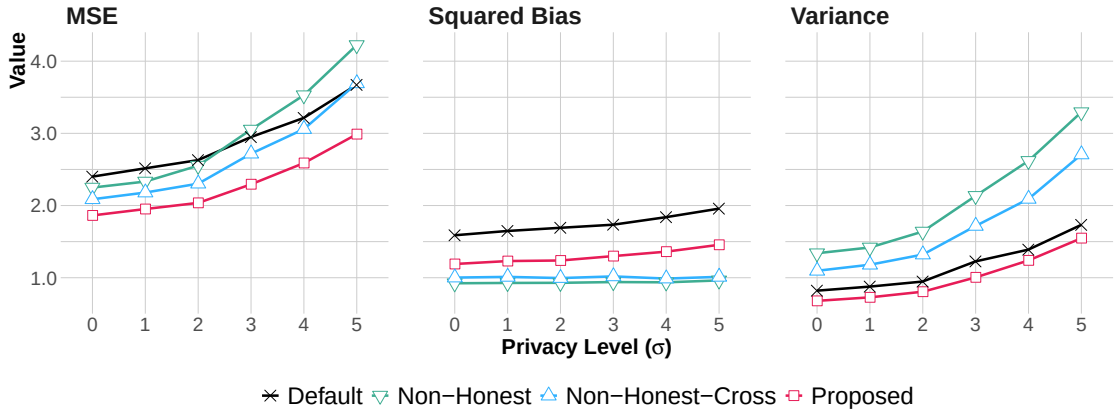
(b) ***Scenario 2: LDP-Protected Outcome***

Note: We simulate 100 replications to compute the bootstrap mean MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point. The results presented here are based on the use of R-learner with regression forests as the initial CATE model.

Figure B.8: *Statistical Accuracy with Varying Privacy Levels: Alternative Post-processing Methods*



(a) ***Scenario 1: LDP-Protected Covariates***



(b) ***Scenario 2: LDP-Protected Outcome***

Note: We simulate 100 replications to compute the bootstrap mean MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point. The results presented here are based on the use of R-learner with regression forests as the initial CATE model.

Figure B.9: *Statistical Accuracy with Varying Privacy Levels: Subgroup Cross-Learning*

B.5.6 Robustness Checks for Different Initial CATE models

In this section, we further evaluate the performance of two different approaches for the initial CATE model:

1. *Causal Forest* [199]: We implement causal forest using the `grf` package with default parameters.

2. *T-learner with regression forests* [134]: We estimate two separate conditional mean outcome models —one for the treatment group and one for the control group—and compute the predicted CATEs as the difference between their predicted outcomes. Each regression model is estimated using regression forests with default parameters from the `grf` package.

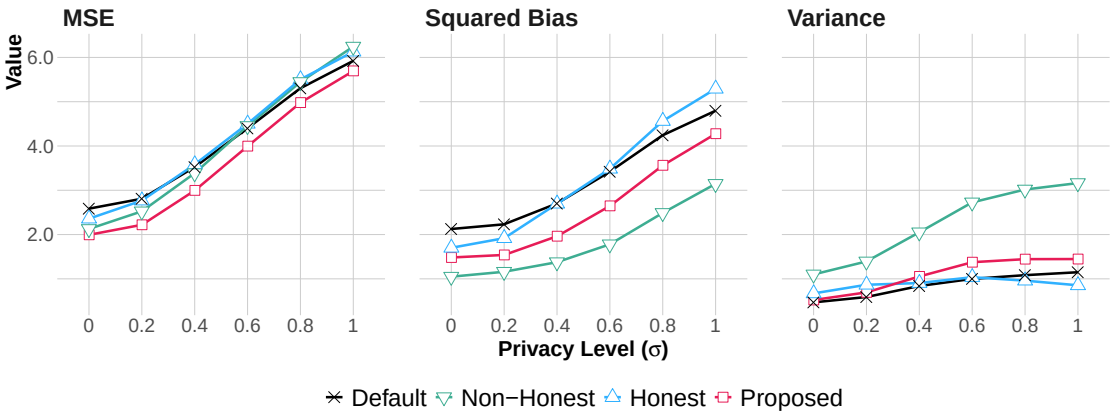
3. *R-learner with XGBoost models* [155]: Similar to the R-learner with regression forests used in the main analysis, we implement R-learner with XGBoost models for both the conditional mean outcome and CATE models. Hyperparameter tuning follows the methods implemented by [155].

4. *T-learner with XGBoost models*: Similar to the T-learner with regression forests, we implement R-learner with XGBoost models for the conditional mean outcome models. Hyperparameter tuning follows the methods implemented by [155].

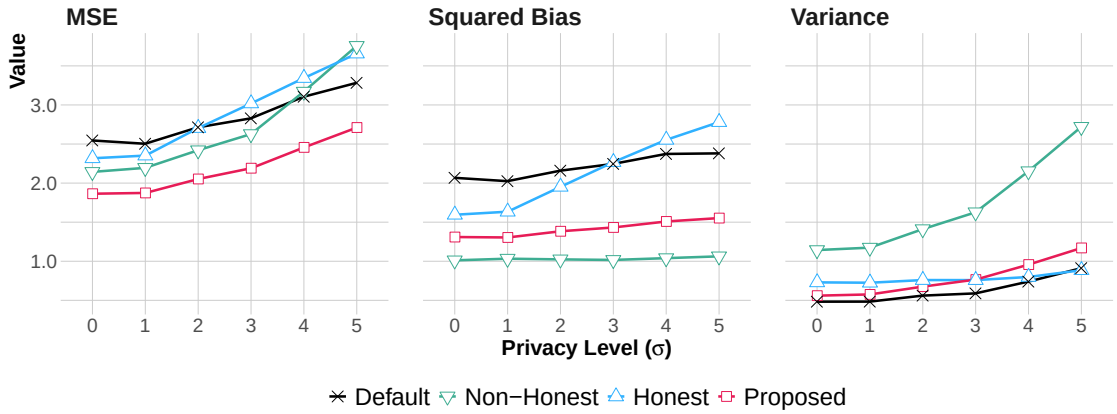
Overall, regardless of the initial CATE models, our proposed solution yields to the smallest predictive error and best targeting performance with varying privacy levels and experiment sample sizes.

Results for Causal Forest.

Varying Privacy Levels. Figure B.10 presents the MSE, squared bias, and variance of different methods across different privacy levels, where we apply Casual Forest as the DEFAULT method and the initial CATE models. Table B.12 reports the AUTOC values across different privacy levels. Table B.13 presents the value improvement of different methods across various privacy levels. Overall, the results align with the findings from the main analysis in Section 2.5.5.



(a) ***Scenario 1: LDP-Protected Covariates***



(b) ***Scenario 2: LDP-Protected Outcome***

Note: We simulate 100 replications to compute the bootstrap mean MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point.

Figure B.10: *Statistical Accuracy with Varying Privacy Levels: Causal Forest*

Table B.12: *AUTOC Values for Different Methods Across Varying Privacy Levels: Causal Forest*

(a) Scenario 1: LDP-Protected Covariates

σ	Proposed	Split-only	Non-Honest	Default
0.0	2.38 (82%)	2.34 (56%)	2.35 (54%)	2.35
0.2	2.33 (94%)	2.28 (70%)	2.28 (56%)	2.28
0.4	2.19 (90%)	2.15 (66%)	2.13 (50%)	2.13
0.6	2.00 (82%)	1.98 (68%)	1.94 (32%)	1.96
0.8	1.81 (80%)	1.79 (72%)	1.73 (28%)	1.77
1.0	1.63 (82%)	1.63 (82%)	1.55 (24%)	1.60

(b) Scenario 2: LDP-Protected Outcome

σ	Proposed	Split-only	Non-Honest	Default
0	2.38 (78%)	2.34 (56%)	2.34 (45%)	2.34
1	2.38 (78%)	2.35 (50%)	2.34 (37%)	2.35
2	2.36 (84%)	2.32 (54%)	2.31 (38%)	2.32
3	2.34 (69%)	2.31 (49%)	2.29 (26%)	2.32
4	2.31 (77%)	2.27 (59%)	2.22 (31%)	2.26
5	2.26 (74%)	2.23 (63%)	2.17 (23%)	2.22

Note: We calculate the AUTOC values from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses).

Table B.13: *Value Improvement Across Varying Privacy Levels: Causal Forest*

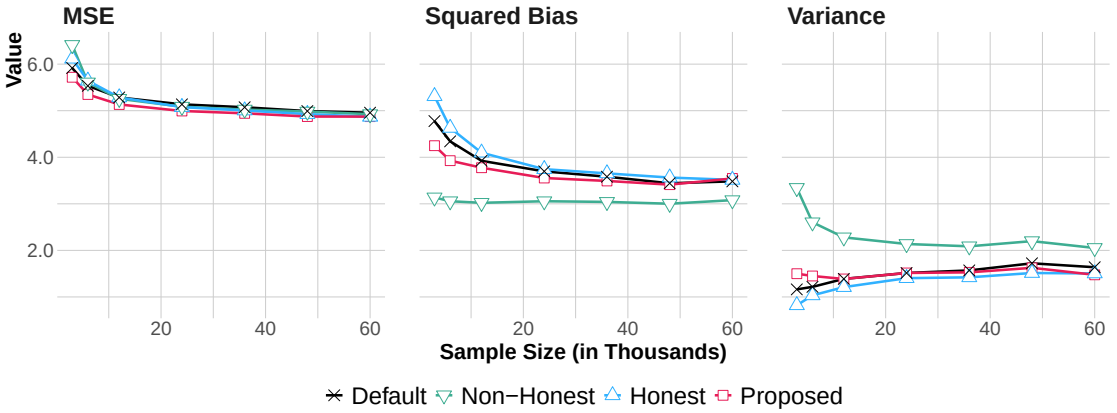
(a) Scenario 1: LDP-Protected Covariates

σ	Proposed	Split-only	Non-Honest
0	10.1% (96%)	3.6% (68%)	8.9% (84%)
0.2	10.4% (98%)	1.9% (64%)	6.4% (78%)
0.4	12.7% (98%)	−3.4% (48%)	7.9% (86%)
0.6	13.4% (84%)	−3.4% (42%)	5.3% (56%)
0.8	14.9% (86%)	−15.9% (24%)	11% (60%)
1	16.8% (92%)	−26.4% (22%)	4.4% (58%)

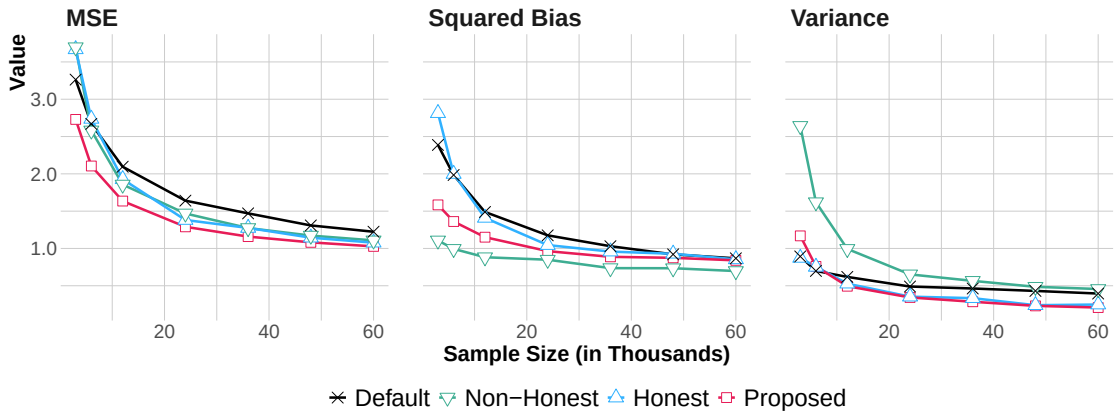
(b) Scenario 2: LDP-Protected Outcome

σ	Proposed	Split-only	Non-Honest
0	9.7% (97%)	2.8% (72%)	5.6% (76%)
1	9.1% (92%)	1.8% (61%)	5.3% (72%)
2	10.6% (93%)	−1.7% (52%)	6.6% (73%)
3	12% (90%)	−5.1% (41%)	8.4% (76%)
4	13.8% (92%)	−6.9% (36%)	7.1% (71%)
5	11.5% (86%)	−11.1% (29%)	2.5% (53%)

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses).



(a) ***Scenario 1: LDP-Protected Covariates***



(b) ***Scenario 2: LDP-Protected Outcome***

Note: We simulate 100 replications to compute the bootstrap mean MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point.

Figure B.11: *Statistical Accuracy with Varying Sample Sizes: Causal Forest*

Varying Sample Sizes. Figure B.11 presents the MSE, squared bias, and variance of different methods across different sample sizes, where we apply Casual Forest as the DEFAULT method and the initial CATE models. Table B.14 reports the AUTOC values across different sample sizes. Table B.15 presents the value improvement of different methods across various sample sizes. Overall, the results align with the findings from the main analysis in Section 2.5.5.

Table B.14: *AUTOC Values for Different Methods Across Varying Sample Sizes: Causal Forest*

(a) Scenario 1: LDP-Protected Covariates

Sample Size	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
3k	1.63 (82%)	1.63 (82%)	1.55 (24%)	1.60
6k	1.69 (92%)	1.69 (92%)	1.64 (30%)	1.58
12k	1.74 (98%)	1.73 (98%)	1.7 (50%)	1.65
24k	1.75 (100%)	1.75 (100%)	1.73 (70%)	1.68
36k	1.76 (100%)	1.76 (100%)	1.74 (85%)	1.70
48k	1.77 (100%)	1.77 (100%)	1.75 (90%)	1.70
60k	1.78 (100%)	1.78 (100%)	1.76 (90%)	1.71

(b) Scenario 2: LDP-Protected Outcome

Sample Size	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
3k	2.26 (74%)	2.23 (63%)	2.17 (23%)	2.22
6k	2.35 (92%)	2.31 (58%)	2.29 (40%)	2.30
12k	2.42 (94%)	2.4 (80%)	2.38 (74%)	2.37
24k	2.45 (100%)	2.45 (95%)	2.43 (85%)	2.42
36k	2.48 (92%)	2.47 (90%)	2.46 (70%)	2.45
48k	2.49 (98%)	2.48 (100%)	2.47 (100%)	2.47
60k	2.50 (96%)	2.49 (80%)	2.49 (90%)	2.48

Note: We calculate the AUTOC values from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses).

Table B.15: *Value Improvement for Different Methods Across Varying Sample Sizes: Causal Forest*

(a) Scenario 1: LDP-Protected Covariates

Sample Size	PROPOSED	SPLIT-ONLY	NON-HONEST
3k	16.8% (92%)	−26.4% (22%)	4.4% (58%)
6k	12.9% (90%)	−12% (26%)	5.5% (58%)
12k	5.1% (94%)	−1.6% (56%)	1.4% (58%)
24k	5.8% (95%)	2.5% (74%)	1.5% (60%)
36k	5.7% (100%)	3.1% (83%)	3% (85%)
48k	3.5% (96%)	2% (87%)	0.3% (52%)
60k	2.5% (90%)	2.9% (100%)	0.5% (61%)

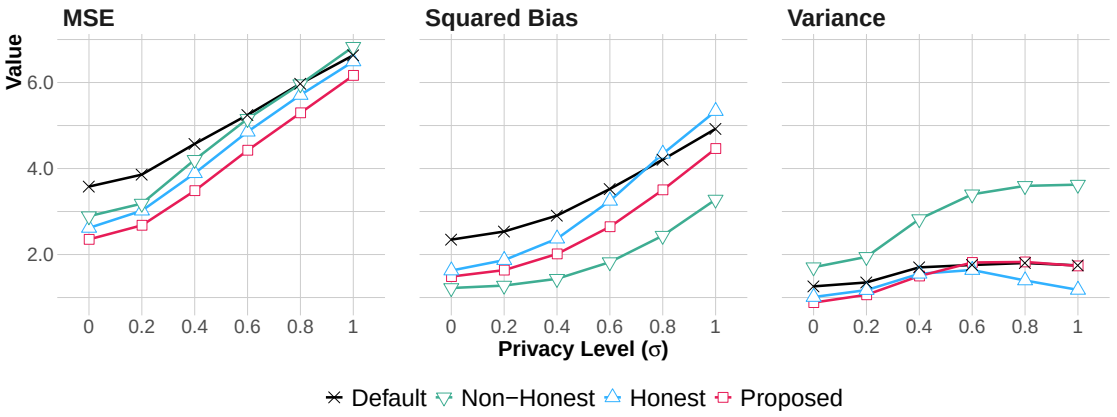
(b) Scenario 2: LDP-Protected Outcome

Sample Size	PROPOSED	SPLIT-ONLY	NON-HONEST
3k	10.1% (82%)	−11.6% (28%)	3.0% (52%)
6k	8.2% (98%)	−0.1% (52%)	2.6% (66%)
12k	6.6% (100%)	2.9% (84%)	3.8% (92%)
24k	3.6% (100%)	2.8% (95%)	1.8% (95%)
36k	3.3% (95%)	2.0% (90%)	2.1% (90%)
48k	1.8% (93%)	1.7% (90%)	0.9% (90%)
60k	1.1% (88%)	0.7% (85%)	0.6% (82%)

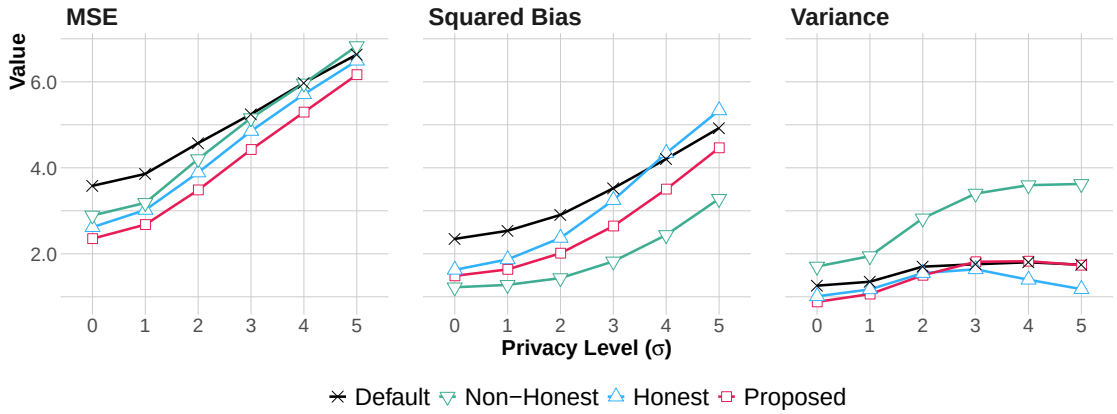
Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses).

Results for T-learner with Regression Forests.

Varying Privacy Levels. Figure B.12 presents the MSE, squared bias, and variance of different methods across different privacy levels, where we apply T-learner with regression forests as the DEFAULT method and the initial CATE models. Table B.16 reports the AUTOC values across different privacy levels. Table B.17 presents the value improvement of different methods across various privacy levels. Overall, the results align with the findings from the main analysis in Section 2.5.5.



(a) ***Scenario 1: LDP-Protected Covariates***



(b) ***Scenario 2: LDP-Protected Outcome***

Note: We simulate 100 replications to compute the bootstrap mean MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point. The results presented here are based on the use of T-learner with regression forests as the initial CATE model.

Figure B.12: *Statistical Accuracy with Varying Privacy Levels: T-learner with Regression Forests*

Table B.16: *AUTOC Values for Different Methods Across Varying Privacy Levels: T-learner with Regression Forests*

(a) Scenario 1: LDP-Protected Covariates

σ	Proposed	Split-only	Non-Honest	Default
0	2.32 (100%)	2.28 (98%)	2.26 (100%)	2.09
0.2	2.26 (100%)	2.21 (100%)	2.2 (100%)	2.03
0.4	2.11 (100%)	2.05 (100%)	2.05 (100%)	1.89
0.6	1.92 (100%)	1.85 (97%)	1.85 (96%)	1.73
0.8	1.72 (100%)	1.66 (92%)	1.68 (96%)	1.56
1	1.50 (98%)	1.46 (92%)	1.46 (96%)	1.37

(b) Scenario 2: LDP-Protected Outcome

σ	Proposed	Split-only	Non-Honest	Default
0	2.28 (100%)	2.24 (98%)	2.22 (100%)	2.04
1	2.27 (100%)	2.23 (100%)	2.21 (100%)	2.05
2	2.24 (100%)	2.18 (99%)	2.19 (100%)	2.02
3	2.22 (100%)	2.16 (100%)	2.16 (99%)	1.99
4	2.17 (100%)	2.08 (96%)	2.10 (100%)	1.93
5	2.10 (97%)	2.01 (91%)	2.05 (98%)	1.88

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses). The results presented here are based on the use of T-learner with regression forests as the initial CATE model.

Table B.17: *Value Improvement Across Varying Privacy Levels: T-learner with Regression Forests*

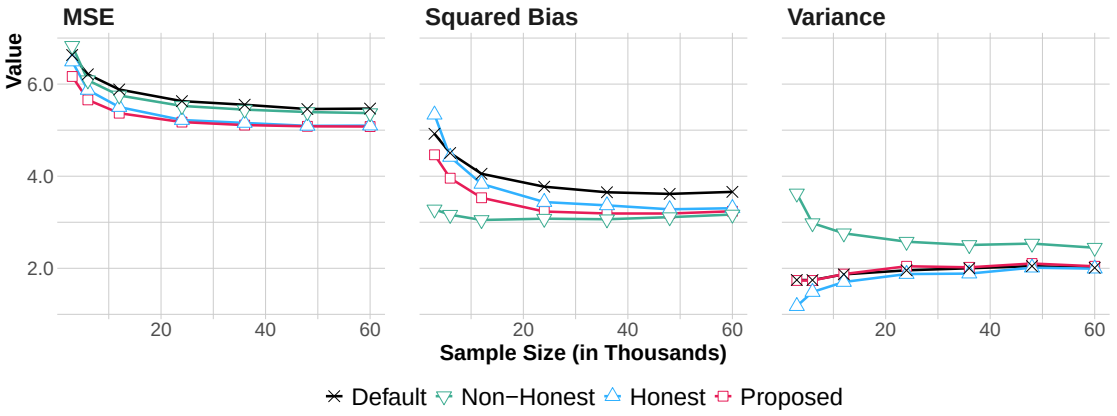
(a) Scenario 1: LDP-Protected Covariates

σ	Proposed	Split-only	Non-Honest
0	22% (100%)	18.4% (98%)	14.2% (98%)
0.2	24.4% (100%)	15.8% (88%)	17.1% (98%)
0.4	30.3% (100%)	19.2% (88%)	18.7% (92%)
0.6	28.9% (100%)	9.9% (80%)	16.0% (84%)
0.8	35.3% (96%)	8.5% (64%)	23.4% (84%)
1	37.3% (94%)	3.4% (54%)	9.8% (62%)

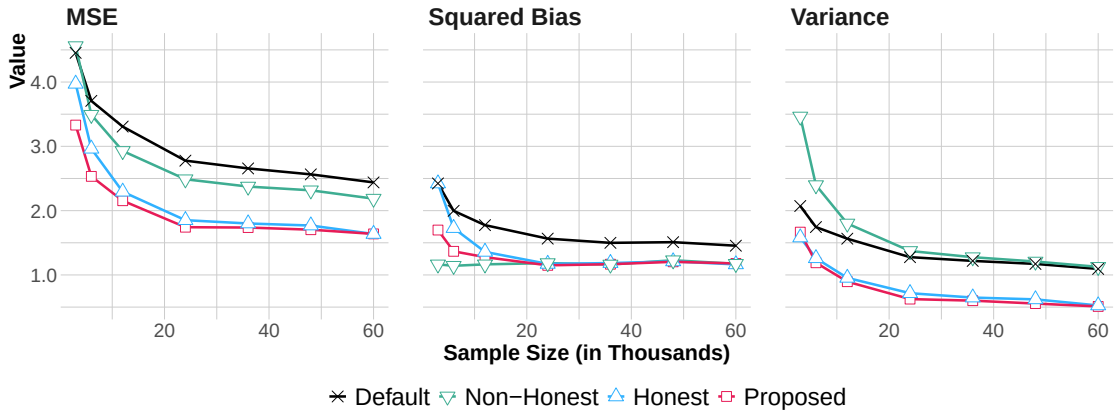
(b) Scenario 2: LDP-Protected Outcome

σ	Proposed	Split-only	Non-Honest
0	25.0% (100%)	18.4% (94%)	17.0% (97%)
1	23.9% (100%)	17.1% (95%)	16.6% (100%)
2	24.9% (100%)	14.0% (91%)	17.6% (97%)
3	26.5% (100%)	15.0% (92%)	18.0% (96%)
4	29.9% (100%)	12.2% (75%)	18.9% (92%)
5	28.7% (97%)	10.2% (73%)	18.5% (86%)

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses). The results presented here are based on the use of T-learner with regression forests as the initial CATE model.



(a) ***Scenario 1: LDP-Protected Covariates***



(b) ***Scenario 2: LDP-Protected Outcome***

Note: We simulate 100 replications to compute the bootstrap mean MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point. The results presented here are based on the use of T-learner with regression forests as the initial CATE model.

Figure B.13: *Statistical Accuracy with Varying Sample Sizes: T-learner with Regression Forests*

Varying Sample Sizes. Figure B.13 presents the MSE, squared bias, and variance of different methods across different sample sizes, where we apply T-learner with regression forests as the DEFAULT method and the initial CATE models.

Table B.18 reports the AUTOC values across different sample sizes. Table B.19 presents the value improvement of different methods across various sample sizes. Overall, the results align with the findings from the main analysis in Section 2.5.5.

Table B.18: *AUTOC Values for Different Methods Across Varying Sample Sizes: T-learner with Regression Forests*

(a) Scenario 1: LDP-Protected Covariates

Sample Size	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
3k	1.5 (98%)	1.46 (92%)	1.46 (96%)	1.37
6k	1.64 (100%)	1.61 (100%)	1.58 (100%)	1.49
12k	1.69 (100%)	1.66 (100%)	1.63 (100%)	1.56
24k	1.72 (100%)	1.72 (100%)	1.67 (100%)	1.62
36k	1.74 (100%)	1.72 (100%)	1.68 (100%)	1.64
48k	1.74 (100%)	1.74 (100%)	1.69 (100%)	1.66
60k	1.74 (100%)	1.74 (100%)	1.69 (100%)	1.65

(b) Scenario 2: LDP-Protected Outcome

Sample Size	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
3k	2.12 (100%)	2.02 (86%)	2.06 (98%)	1.89
6k	2.25 (100%)	2.18 (100%)	2.17 (100%)	2.02
12k	2.30 (100%)	2.28 (100%)	2.22 (100%)	2.10
24k	2.36 (100%)	2.34 (100%)	2.28 (100%)	2.19
36k	2.37 (100%)	2.35 (100%)	2.29 (100%)	2.22
48k	2.37 (100%)	2.36 (100%)	2.30 (100%)	2.23
60k	2.38 (100%)	2.38 (100%)	2.31 (100%)	2.26

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses). The results presented here are based on the use of T-learner with regression forests as the initial CATE model.

Table B.19: *Value Improvement for Different Methods Across Varying Sample Sizes: T-learner with Regression Forests*

(a) Scenario 1: LDP-Protected Covariates

Sample Size	PROPOSED	SPLIT-ONLY	NON-HONEST
3k	37.3% (94%)	3.4% (54%)	9.8% (62%)
6k	35.1% (100%)	19.9% (88%)	16.9% (84%)
12k	25.2% (100%)	18.3% (98%)	9.6% (94%)
24k	17.5% (100%)	16.0% (100%)	4.0% (92%)
36k	16.5% (100%)	15.0% (100%)	3.0% (94%)
48k	13.5% (100%)	12.5% (100%)	3.3% (95%)
60k	15.7% (100%)	14.7% (100%)	3.6% (95%)

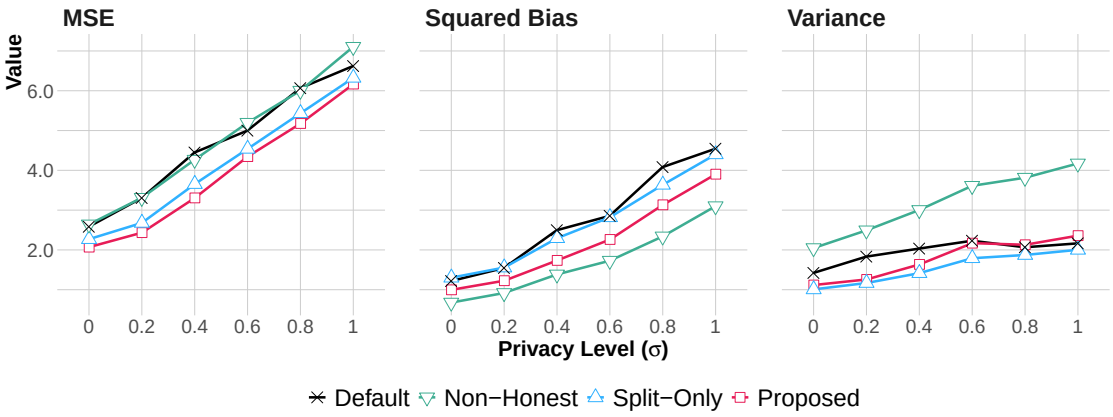
(b) Scenario 2: LDP-Protected Outcome

Sample Size	PROPOSED	SPLIT-ONLY	NON-HONEST
3k	29.4% (98%)	10.9% (80%)	18.2% (82%)
6k	23.2% (100%)	14.9% (100%)	11.8% (94%)
12k	18.2% (100%)	15.2% (98%)	9.1% (100%)
24k	14.2% (100%)	12.2% (100%)	5.8% (100%)
36k	11.9% (100%)	11% (100%)	4.1% (100%)
48k	10.4% (100%)	9.7% (100%)	3.8% (100%)
60k	9.9% (100%)	9.7% (100%)	3.4% (100%)

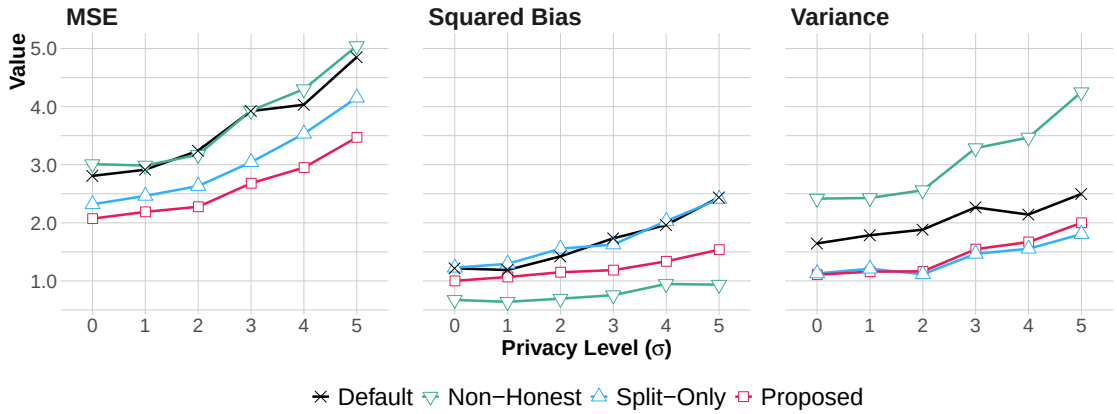
Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses). The results presented here are based on the use of T-learner with regression forests as the initial CATE model.

Results for R-learners with XGBoost Models.

Varying Privacy Levels. Figure B.14 presents the MSE, squared bias, and variance of different methods across different privacy levels, where we apply R-learner with XGBoost models as the DEFAULT method and the initial CATE models.



(a) ***Scenario 1: LDP-Protected Covariates***



(b) ***Scenario 2: LDP-Protected Outcome***

Note: We simulate 100 replications to compute the bootstrap mean MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point. The results presented here are based on the use of R-learner with XGBoost models as the initial CATE model.

Figure B.14: *Statistical Accuracy with Varying Privacy Levels: R-learner with XGBoost Models*

Table B.20 reports the AUTOC values across different privacy levels. Table B.21 presents the value improvement of different methods across various privacy levels. Overall, the results align with the findings from the main analysis in Section 2.5.5.

Table B.20: *AUTOC Values for Different Methods Across Varying Sample Sizes: R-learner with XGBoost Models*

(a) Scenario 1: LDP-Protected Covariates

σ	Proposed	Split-only	Non-Honest	Default
0	2.32 (72%)	2.31 (72%)	2.27 (48%)	2.27
0.2	2.26 (100%)	2.23 (84%)	2.17 (72%)	2.13
0.4	2.10 (96%)	2.07 (92%)	2.00 (88%)	1.93
0.6	1.91 (88%)	1.89 (86%)	1.84 (68%)	1.81
0.8	1.72 (88%)	1.69 (88%)	1.65 (88%)	1.56
1	1.49 (80%)	1.48 (80%)	1.42 (72%)	1.38

(b) Scenario 2: LDP-Protected Outcome

σ	Proposed	Split-only	Non-Honest	Default
0	2.36 (93%)	2.33 (93%)	2.27 (67%)	2.25
1	2.35 (87%)	2.33 (83%)	2.28 (67%)	2.24
2	2.33 (97%)	2.29 (83%)	2.24 (83%)	2.18
3	2.27 (97%)	2.25 (97%)	2.18 (87%)	2.10
4	2.22 (90%)	2.17 (70%)	2.13 (73%)	2.06
5	2.16 (90%)	2.10 (83%)	2.07 (87%)	1.96

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses). The results presented here are based on the use of R-learner with XGBoost models as the initial CATE model.

Table B.21: *Value Improvement Across Varying Privacy Levels: R-learner with XGBoost Models*

(a) Scenario 1: LDP-Protected Covariates

σ	Proposed	Split-only	Non-Honest
0	4.8% (72%)	4.1% (64%)	0.3% (40%)
0.2	14.1% (96%)	11.7% (84%)	4.1% (73%)
0.4	25.0% (98%)	22% (88%)	11% (92%)
0.6	18.2% (80%)	16% (76%)	4% (60%)
0.8	36.8% (88%)	28.7% (85%)	3.9% (63%)
1	40.3% (76%)	37.7% (79%)	−0.2% (36%)

(b) Scenario 2: LDP-Protected Outcome

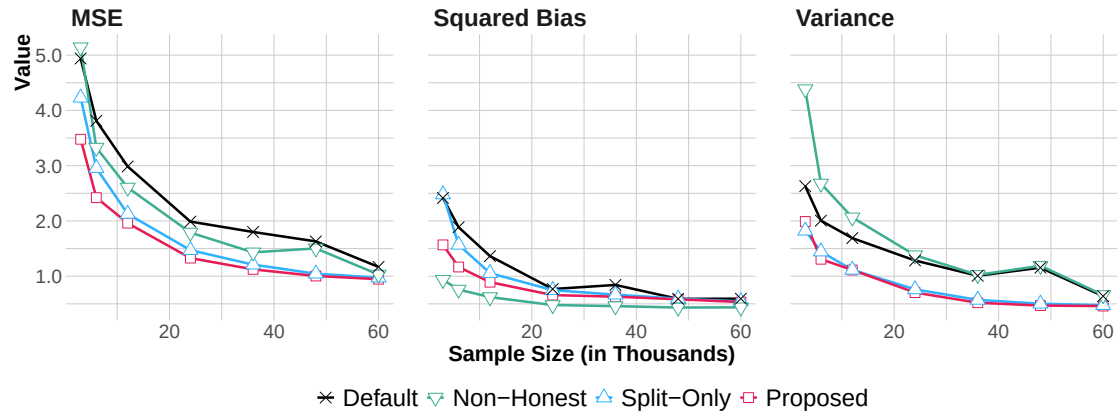
σ	Proposed	Split-only	Non-Honest
0	10.0% (93%)	7.3% (90%)	1.7% (63%)
1	9.3% (77%)	7.7% (70%)	2.5% (67%)
2	14.3% (90%)	10.5% (80%)	5.3% (70%)
3	21.4% (97%)	18.6% (97%)	9.1% (73%)
4	20.1% (95%)	13.1% (73%)	7.7% (73%)
5	35.7% (93%)	26.9% (82%)	18.6% (83%)

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses). The results presented here are based on the use of R-learner with XGBoost models as the initial CATE model.

Varying Sample Sizes. Figure B.15 presents the MSE, squared bias, and variance of different methods across different sample sizes, where we apply R-learner with XGBoost models as the Default method and the initial CATE models.

Table B.22 reports the AUTOC values across different sample sizes. Table B.23 presents the value improvement of different methods across various sample sizes. Overall, the results align with the findings from the main analysis in Section 2.5.5.

(a) ***Scenario 1: LDP-Protected Covariates***



(b) ***Scenario 2: LDP-Protected Outcome***

Note: We simulate 100 replications to compute the bootstrap mean MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point. The results presented here are based on the use of R-learner with XGBoost models as the initial CATE model.

Figure B.15: *Statistical Accuracy with Varying Sample Sizes: R-learner with XGBoost Models*

Table B.22: *AUTOC Values for Different Methods Across Varying Sample Sizes: R-learner with XGBoost Modelss*

(a) Scenario 1: LDP-Protected Covariates

Sample Size	Proposed	Split-only	Non-Honest	Default
3k	1.49 (80%)	1.48 (80%)	1.42 (72%)	1.38
6k	1.59 (92%)	1.57 (82%)	1.53 (88%)	1.46
12k	1.66 (92%)	1.64 (86%)	1.6 (84%)	1.54
24k	1.71 (83%)	1.7 (70%)	1.68 (90%)	1.66
36k	1.73 (81%)	1.72 (80%)	1.70 (80%)	1.68
48k	1.73 (85%)	1.73 (77%)	1.73 (83%)	1.72
60k	1.74 (94%)	1.74 (87%)	1.72 (88%)	1.70

(b) Scenario 2: LDP-Protected Outcome

Sample Size	Proposed	Split-only	Non-Honest	Default
3k	2.16 (90%)	2.10 (83%)	2.07 (87%)	1.96
6k	2.30 (84%)	2.25 (72%)	2.25 (72%)	2.11
12k	2.37 (96%)	2.36 (96%)	2.36 (96%)	2.23
24k	2.46 (75%)	2.45 (74%)	2.45 (69%)	2.37
36k	2.49 (81%)	2.48 (75%)	2.48 (75%)	2.42
48k	2.51 (93%)	2.50 (85%)	2.50 (83%)	2.43
60k	2.52 (80%)	2.52 (77%)	2.52 (60%)	2.50

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses). The results presented here are based on the use of R-learner with XGBoost models as the initial CATE model.

Table B.23: *Value Improvement for Different Methods Across Varying Sample Sizes: R-learner with XGBoost Models*

(a) Scenario 1: LDP-Protected Covariates

Sample Size	Proposed	Split-only	Non-Honest
3k	40.3% (76%)	37.7% (79%)	−0.2% (36%)
6k	51.6% (82%)	47.5% (84%)	23.8% (84%)
12k	27.6% (88%)	26.7% (84%)	11.7% (84%)
24k	11% (85%)	9.5% (80%)	4.2% (75%)
36k	9.0% (86%)	8.1% (78%)	2.5% (78%)
48k	8.5% (83%)	7.6% (76%)	1.4% (71%)
60k	7.7% (90%)	7.4% (90%)	4.8% (100%)

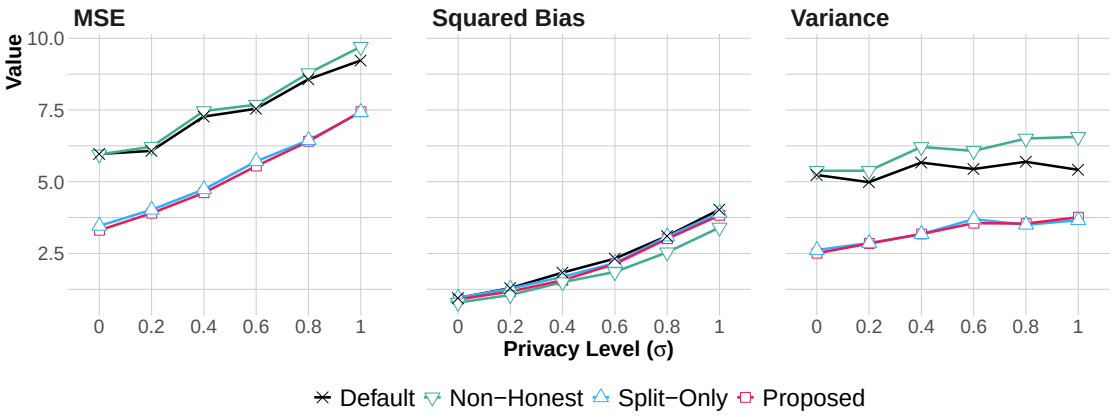
(b) Scenario 2: LDP-Protected Outcome

Sample Size	Proposed	Split-only	Non-Honest
3k	35.7% (93%)	26.9% (82%)	18.6% (83%)
6k	24.2% (96%)	17.3% (68%)	14.6% (80%)
12k	13.8% (88%)	12.6% (92%)	7.1% (84%)
24k	7.3% (86%)	6.2% (65%)	3.1% (85%)
36k	5% (83%)	4.3% (80%)	2.4% (75%)
48k	7.6% (80%)	7.1% (77%)	2.6% (90%)
60k	1.5% (75%)	1.30% (69%)	0.8% (60%)

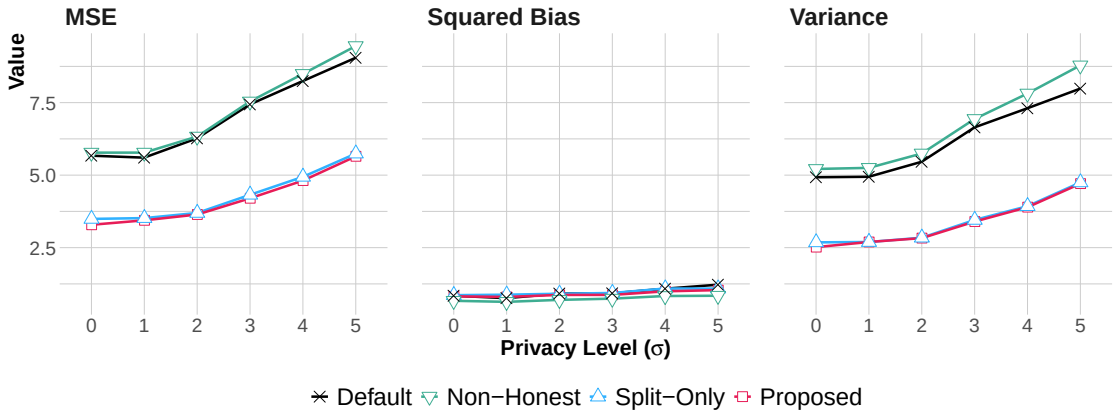
Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses). The results presented here are based on the use of R-learner with XGBoost models as the initial CATE model.

Results for T-learners with XGBoost Models.

Varying Privacy Levels. Figure B.16 presents the MSE, squared bias, and variance of different methods across different privacy levels, where we apply T-learners with XGBoost models as the Default method and the initial CATE models. Table B.12 reports the AUTOC values across different privacy levels. Table B.13 presents the value improvement of different methods across various privacy levels. Overall, the results align with the findings from the main analysis in Section 2.5.5.



(a) ***Scenario 1: LDP-Protected Covariates***



(b) ***Scenario 2: LDP-Protected Outcome***

Note: We simulate 100 replications to compute the bootstrap mean MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point. The results presented here are based on the use of T-learner with XGBoost models as the initial CATE model.

Figure B.16: *Statistical Accuracy with Varying Privacy Levels: T-learner with XGBoost Models*

Table B.24: *AUTOC Values for Different Methods Across Varying Privacy Levels: T-learner with XGBoost Models*

(a) Scenario 1: LDP-Protected Covariates

σ	Proposed	Split-only	Non-Honest	Default
0	2.18 (100%)	2.16 (100%)	1.98 (88%)	1.91
0.2	2.09 (100%)	2.05 (97%)	1.93 (100%)	1.85
0.4	1.95 (100%)	1.92 (100%)	1.73 (80%)	1.66
0.6	1.76 (96%)	1.74 (96%)	1.63 (92%)	1.54
0.8	1.52 (94%)	1.51 (93%)	1.39 (95%)	1.30
1	1.31 (100%)	1.3 (100%)	1.19 (88%)	1.12

(b) Scenario 2: LDP-Protected Outcome

σ	Proposed	Split-only	Non-Honest	Default
0	2.20 (100%)	2.17 (100%)	2.01 (90%)	1.96
1	2.18 (99%)	2.15 (94%)	2.02 (84%)	1.98
2	2.14 (98%)	2.13 (95%)	1.97 (89%)	1.89
3	2.08 (97%)	2.06 (98%)	1.9 (91%)	1.81
4	2.00 (98%)	1.97 (96%)	1.83 (98%)	1.73
5	1.91 (95%)	1.89 (95%)	1.78 (96%)	1.64

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses). The results presented here are based on the use of T-learner with XGBoost models as the initial CATE model.

Table B.25: *Value Improvement Across Varying Privacy Levels: T-learner with XGBoost Models*

(a) Scenario 1: LDP-Protected Covariates

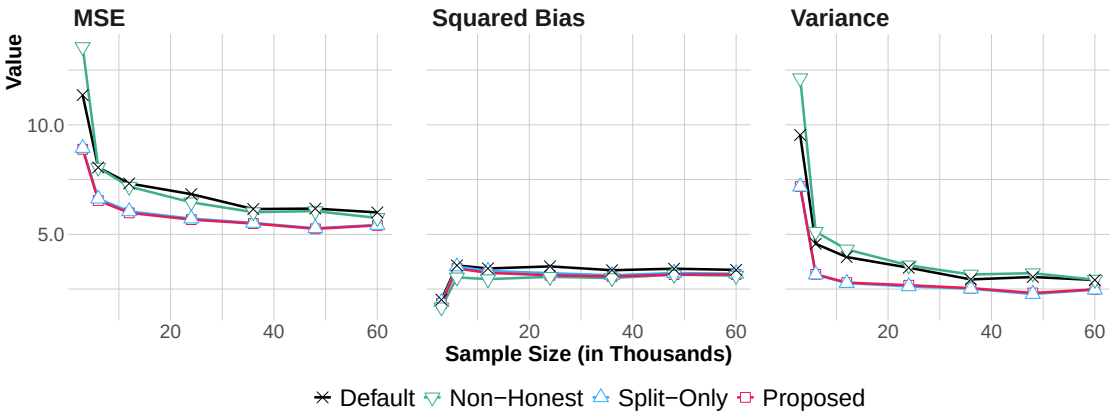
σ	Proposed	Split-only	Non-Honest
0	46.8% (100%)	42.6% (100%)	8.4% (76%)
0.2	46.8% (100%)	41.3% (100%)	13.7% (92%)
0.4	85% (100%)	78.1% (100%)	14% (64%)
0.6	82.4% (92%)	73.6% (92%)	23% (72%)
0.8	664.3% (96%)	650.4% (92%)	141.9% (72%)
1	540.1% (92%)	508% (96%)	59.1% (92%)

(b) Scenario 2: LDP-Protected Outcome

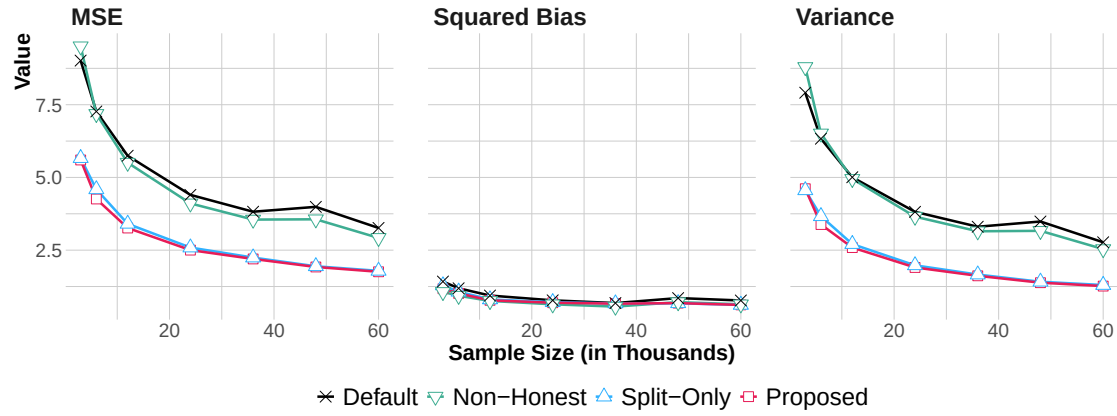
σ	Proposed	Split-only	Non-Honest
0	37.6% (98%)	33.3% (98%)	4.5% (54%)
1	34.2% (100%)	31.0% (96%)	3.8% (64%)
2	46.9% (98%)	44.4% (97%)	9.0% (70%)
3	63.7% (99%)	58.7% (98%)	16.3% (76%)
4	69.8% (95%)	62.8% (94%)	13.6% (70%)
5	122.4% (98%)	116.7% (98%)	47.4% (84%)

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses). The results presented here are based on the use of T-learner with XGBoost models as the initial CATE model.

Varying Sample Sizes. Figure B.17 presents the MSE, squared bias, and variance of different methods across different sample sizes, where we apply T-learner with XGBoost models as the Default method and the initial CATE models. Table B.26 reports the AUTOC values across different sample sizes. Table B.27 presents the value improvement of different methods across various sample sizes. Overall, the results align with the findings from the main analysis in Section 2.5.5.



(a) ***Scenario 1: LDP-Protected Covariates***



(b) ***Scenario 2: LDP-Protected Outcome***

Note: We simulate 100 replications to compute the bootstrap mean MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point. The results presented here are based on the use of T-learner with XGBoost models as the initial CATE model.

Figure B.17: *Statistical Accuracy with Varying Sample Sizes: T-learner with XGBoost Modelss*

Table B.26: *AUTOC Values for Different Methods Across Varying Sample Sizes: T-learner with XGBoost Models*

(a) Scenario 1: LDP-Protected Covariates

Sample Size	Proposed	Split-only	Non-Honest	Default
3k	1.52 (98%)	1.51 (94%)	1.36 (58%)	1.35
6k	1.46 (96%)	1.44 (90%)	1.35 (96%)	1.27
12k	1.56 (92%)	1.54 (94%)	1.45 (98%)	1.38
24k	1.62 (100%)	1.6 (93%)	1.53 (100%)	1.44
36k	1.65 (95%)	1.65 (80%)	1.59 (100%)	1.54
48k	1.69 (100%)	1.68 (100%)	1.58 (100%)	1.54
60k	1.67 (83%)	1.66 (80%)	1.62 (100%)	1.57

(b) Scenario 2: LDP-Protected Outcome

Sample Size	Proposed	Split-only	Non-Honest	Default
3k	1.91 (100%)	1.89 (95%)	1.78 (96%)	1.64
6k	2.06 (100%)	2.02 (96%)	1.92 (100%)	1.8
12k	2.20 (100%)	2.18 (92%)	2.05 (96%)	1.95
24k	2.30 (100%)	2.28 (100%)	2.16 (100%)	2.08
36k	2.33 (100%)	2.33 (100%)	2.21 (100%)	2.15
48k	2.38 (100%)	2.37 (100%)	2.22 (100%)	2.13
60k	2.39 (100%)	2.39 (100%)	2.28 (100%)	2.22

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses). The results presented here are based on the use of T-learner with XGBoost models as the initial CATE model.

Table B.27: *Value Improvement for Different Methods Across Varying Sample Sizes: T-learner with XGBoost Models*

(a) Scenario 1: LDP-Protected Covariates

Sample Size	Proposed	Split-only	Non-Honest
3k	540.1% (92%)	508% (96%)	51.3% (48%)
6k	290.4% (94%)	273.5% (94%)	75.8% (76%)
12k	100.6% (92%)	94.3% (94%)	25.5% (68%)
24k	78.0% (90%)	74.3% (85%)	37.8% (75%)
36k	29.7% (90%)	27.9% (93%)	10.4% (80%)
48k	40.7% (100%)	39.2% (100%)	6.8% (65%)
60k	22.4% (88%)	21.1% (75%)	10.6% (43%)

(b) Scenario 2: LDP-Protected Outcome

Sample Size	Proposed	Split-only	Non-Honest
3k	122.4% (98%)	116.7% (98%)	47.4% (84%)
6k	60.8% (100%)	51.5% (96%)	21.1% (84%)
12k	43.8% (100%)	39.3% (92%)	15.9% (92%)
24k	26.4% (100%)	24.8% (100%)	8.4% (90%)
36k	21.5% (95%)	20.6% (95%)	7% (95%)
48k	29.3% (96%)	28.9% (93%)	9.7% (87%)
60k	16.1% (98%)	15.5% (94%)	3.8% (68%)

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses). The results presented here are based on the use of T-learner with XGBoost models as the initial CATE model.

B.6 Further Details about Hillstrom Case Study

In this section, we provide the summary statistics, implementation details, and robustness checks for the Hillstrom case study in Section 2.6.4.

B.6.1 Summary Statistics

Table B.28 presents the summary statistics for the Hillstrom data. The definitions of the pre-treatment covariates are:

1. Recency: The number of months since the customer's last purchase.
2. History: The actual dollar value the customer has spent in the past year.
3. Mens: A binary variable indicating whether the customer purchased men's merchandise in the past year (1 = Yes).
4. Womens: A binary variable indicating whether the customer purchased women's merchandise in the past year (1 = Yes).
5. Zip_Code: A classification of the customer's zip code as Urban, Suburban, or Rural
6. Newbie: A binary variable indicating whether the customer was acquired in the past twelve months (1 = Yes).
7. Channel: The channel(s) the customer purchased from in the past year.

Table B.28: *Summary Statistics for Hillstrom Data*

Discrete Variables

Variable	N	Unique Values	Distributions
visit (Outcome)	42,693	2	0: 37,193, 1: 5,500
email (Treatment)	42,693	2	0: 21,306, 1: 21,387
mens	42,693	2	0: 19,166, 1: 23,527
womens	42,693	2	0: 19,260, 1: 23,433
newbie	42,693	2	0: 21,235, 1: 21,458
channel	42,693	3	Phone: 18,781, Website: 18,727, Multichannel: 5,185
zip_code	42,693	3	Suburban: 19,275, Urban 17,098, Rural: 6,320

Continuous Variables

Variable	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
history	42,693	241.71	254.04	29.99	65.16	158.46	326.05	3345.93
newbie	42,693	0.5	0.5	0	0	1	1	1
recency	42,693	5.76	3.5	1	2	5	9	12

B.6.2 Covariate Balance Check

To assess covariate balance, we compare the distributions of each covariate for both treated and non-treated customers. In particular, we use the *standardized mean difference* measure, which is the mean difference between the treated and non-treated groups divided by the pooled standard deviation. Generally, it is considered small if the value is less than 0.20 [48]. Table B.29 reports the summary for the covariate balance check. Note that the standardized mean differences are close to zero for all covariates, suggesting that the experiment is properly randomized.

Table B.29: *Covariate Balance Check for Hillstrom Data*

Variable	Mean Diff.	Pooled St. Dev.	Standardized Mean Diff.
recency	0.021	3.505	0.006
history	1.803	255.307	0.007
mens	−0.003	0.497	−0.007
womens	0.003	0.498	0.006
newbie	0.000	0.500	0.001
channel_Multichannel	−0.002	0.327	−0.005
channel_Phone	0.000	0.496	0.000
channel_Web	0.001	0.496	0.003
zip_code_Rural	0.003	0.356	0.009
zip_code_Suburban	−0.003	0.498	−0.006
zip_code_Urban	0.000	0.490	0.000

We also conduct a linear regression test to assess whether any covariate predicts the treatment assignment indicator. The joint F-test indicates that the model has no predictive power (F-stat = 0.4292, p-value = 0.9202), suggesting that randomization was successfully implemented.

B.6.3 Implementation of Local Differential Privacy

We examine two scenarios: the LDP-protected covariates and LDP-protected outcome. In the first scenario, we deploy the Laplace mechanism for the `recency` variable, setting $\sigma_{\texttt{recency}}$ to values in the set $2, 4, 6, 8, 10$, and for the `history` variable, setting $\sigma_{\texttt{recency}}$ to values in the set $20, 40, 60, 80, 100$. For all discrete variables, we implement the randomized response mechanism, with p in the set $0.05, 0.10, 0.15, 0.20, 0.25$. In the secon scenario, we apply the randomized response mechanism described in Section 2.6.2, setting p to values in the set $0.05, 0.10, 0.15, 0.20, 0.25$.

B.6.4 Model Specification

In the main analysis, we use the R-learner with regression forest models models and five-fold cross-fitting for the DEFAULT method and initial CATE models. We use a constant propensity score (0.5) for Robinson's transformation. For all regression forest models, we use 500 trees instead of the default 2,000 trees to accelerate training, while keeping all other parameters at their default settings in the `grf` package.

To construct the AIPW score for post-processing, we set a constant propensity score ($\widehat{e} = 0.5$), considering that the experiment is completely randomized. As for the conditional mean outcome models, we evaluate three candidate models: linear regression, logistic regression, and regression forest. The data is split into two sets: 70% allocated as the training set and the remaining 30% serving as the holdout set. The mean-squared error (MSE) metric is reported in Table B.30, calculated as $\frac{1}{N_{\text{holdout}}} \sum_{l \in \text{holdout set}} [Y_i - \widehat{m}_{W_l}(\widetilde{\mathbf{X}}_l)]^2$. Linear regression is chosen as our final model since it yields the smallest MSE.

Table B.30: *MSE of Conditional Mean Outcome Models*

Privacy	Scenario 1: LDP-Protected Covariates			Scenario 2: LDP-Protected Outcome		
	Linear Regression	Logistic Regression	Regression Forest	Linear Regression	Logistic	Regression Forest
No	0.1088	0.1089	0.1099	0.1088	0.1089	0.1099
Very Low	0.1094	0.1094	0.1098	0.1118	0.1119	0.1128
Low	0.1118	0.1119	0.1123	0.1107	0.1109	0.1122
Medium	0.1092	0.1092	0.1093	0.1184	0.1184	0.1205
High	0.1155	0.1155	0.1159	0.1186	0.1186	0.1204
Very High	0.1108	0.1109	0.1114	0.1202	0.1202	0.1220

For the adjustment models in the post-processing procedures, we use linear regression to ensure simplicity and computational efficiency. We set the number of subgroups to 10 and the

maximum number of iterations to 20 for the subgroup cross-learning algorithm.

B.6.5 Small Sample Performance

Here, we demonstrate that in small-sample settings, the Proposed method outperforms the Split-Only method. Instead of using 70% for model construction and 30% for holdout evaluation, we now only use 10% of the data to construct models and 90% to evaluate the performance.

Table B.31 presents the RMSE across varying privacy levels. The results show that the Proposed method significantly outperforms the Split-Only method, with both improving accuracy compared to the Default approach. Additionally, the Non-Honest post-processing method can even degrade performance, underscoring the importance of our Proposed method in small-sample settings.

Table B.31: *RMSE of GATE (Multiplied by 100) for Different Methods Across Varying Privacy Levels*

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Split-only	Non-Honest	Default	Proposed	Split-only	Non-Honest	Default
No	4.00 (100%)	4.13 (100%)	7.20 (4%)	6.88	4.00 (100%)	4.13 (100%)	7.20 (4%)	6.88
Very Low	4.02 (100%)	4.19 (100%)	6.62 (24%)	6.46	4.26 (100%)	4.38 (100%)	7.55 (16%)	7.24
Low	3.90 (100%)	4.06 (100%)	6.46 (30%)	6.33	4.66 (100%)	4.69 (100%)	8.21 (4%)	7.92
Medium	4.01 (100%)	4.09 (100%)	6.52 (16%)	6.35	5.07 (100%)	5.14 (100%)	8.75 (4%)	8.31
High	3.90 (100%)	4.07 (100%)	6.45 (20%)	6.28	5.23 (100%)	5.36 (100%)	9.06 (4%)	8.74
Very High	3.92 (100%)	4.03 (100%)	6.43 (24%)	6.26	5.21 (100%)	5.41 (100%)	9.19 (0%)	8.84

Note: We calculate the RMSE from 50 random splits, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses).

Table B.32 reports the AUTOC values across different privacy levels. Consistent with previous findings, the Proposed method outperforms the Split-Only method, with both achieving better accuracy than the Default approach. While the Non-Honest post-processing method also enhances treatment prioritization ability, it suffers from higher predictive error.

Table B.32: *AUTOC (Multiplied by 100) for Different Methods Across Varying Privacy Levels*

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Split-only	Non-Honest	Default	Proposed	Split-only	Non-Honest	Default
No	1.08 (96%)	1.01 (92%)	0.87 (96%)	0.70	1.08 (96%)	1.01 (92%)	0.87 (96%)	0.70
Very Low	1.00 (94%)	1.00 (94%)	0.83 (100%)	0.62	0.83 (96%)	0.84 (96%)	0.76 (100%)	0.59
Low	0.86 (96%)	0.85 (94%)	0.72 (100%)	0.54	0.63 (88%)	0.66 (76%)	0.64 (92%)	0.45
Medium	0.83 (84%)	0.76 (80%)	0.77 (98%)	0.59	0.55 (80%)	0.51 (64%)	0.50 (96%)	0.34
High	0.78 (90%)	0.74 (82%)	0.68 (84%)	0.49	0.56 (82%)	0.52 (80%)	0.53 (82%)	0.40
Very High	0.72 (90%)	0.67 (76%)	0.64 (86%)	0.47	0.50 (86%)	0.49 (72%)	0.45 (88%)	0.25

Note: We calculate the AUTOC values from 50 random splits, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses).

B.6.6 Value Improvement for Targeting

In addition to treatment prioritization, we evaluate the improvement in targeting performance achieved by our proposed method when decision-makers rely on CATE models estimated from LDP-protected data. Following B.5.2, we quantify the policy value gained by targeting individuals with positive predicted CATEs using the inverse probability-weighting estimator [217]. Here, the propensity score is estimated as the empirical probability that the targeting policy aligns with the treatment assignment in the data. We then report the value improvement of a method $\widehat{\tau}$ compared to the DEFAULT method $\widehat{\tau}^{\text{DEFAULT}}$ as the percentage difference in their targeting value, i.e., $100\% \times \frac{V(\widehat{\tau}) - V(\widehat{\tau}^{\text{DEFAULT}})}{V(\widehat{\tau}^{\text{DEFAULT}})}$.

Table B.33 reports the average value improvement across different privacy levels. The results indicate that both the PROPOSED method and the SPLIT-ONLY method consistently outperform the default approach, with the proposed method achieving slightly better performance than the split-only method.

Table B.33: *Value Improvement Across Varying Privacy Levels: R-learners with Regression Forests*

Privacy	Scenario 1: LDP-Protected Covariates			Scenario 2: LDP-Protected Outcome		
	PROPOSED	SPLIT-ONLY	NON-HONEST	PROPOSED	SPLIT-ONLY	NON-HONEST
No	2.25% (88%)	2.22% (92%)	−0.03% (56%)	2.25% (88%)	2.22% (92%)	−0.03% (56%)
Very Low	2.25% (92%)	2.13% (86%)	0.10% (50%)	2.32% (94%)	2.33% (98%)	0.06% (58%)
Low	2.28% (90%)	2.36% (92%)	0.03% (48%)	2.48% (92%)	2.51% (94%)	0.24% (66%)
Medium	2.73% (94%)	2.59% (92%)	0.10% (56%)	2.73% (98%)	2.64% (94%)	0.21% (70%)
High	2.64% (96%)	2.58% (96%)	0.12% (60%)	2.85% (94%)	2.85% (98%)	0.23% (72%)
Very High	2.88% (96%)	2.84% (98%)	0.16% (64%)	3.24% (100%)	3.18% (98%)	0.16% (66%)

Note: We calculate the value improvement from 50 splits, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses). The results presented here are based on the use of R-learner with regression forests as the initial CATE model. Results derived from different CATE models are available in Electronic Companion B.5.6.

B.6.7 Robustness Checks of Other Default Models

Forest-based Models.

We begin by considering two alternative forest-based CATE models as the DEFAULT method and the inital CATE models for the post-processing algorithms: Causal Forest and T-learner with regression forests.

1. *Causal Forest*: We use the causal forest function implemented in the `grf` package with 500 trees and other default parameters. Note that we choose 500 trees instead the default 2,000 trees to accelerate the model training process.

2. *T-learner*: We estimate two separate conditional mean outcome models — one for the treatment group and one for the control group—and compute the predicted CATEs as the difference between their predicted outcomes. For each regression forest, we use 500 trees instead of the default 2,000 trees to accelerate the training process, while keeping all other parameters at their default settings in the `grf` package.

Table B.34 reports the RMSE for the two CATE models. The results indicate that (i) the Proposed method significantly outperforms all other methods, (ii) the Split-Only method also improves upon the Default method, while (iii) the Non-Honest post-processing method fail to enhance predictive performance.

Table B.34: *RMSE of GATE (Multiplied by 100) for Different Methods Across Varying Privacy Levels*

(a) T-learners with Regression Forests

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Split-only	Non-Honest	Default	Proposed	Split-only	Non-Honest	Default
No	2.51 (96%)	2.69 (96%)	3.66 (60%)	3.70	2.51 (96%)	2.69 (96%)	3.66 (60%)	3.70
Very Low	3.01 (96%)	3.09 (96%)	3.69 (48%)	3.73	2.41 (100%)	2.46 (100%)	3.88 (36%)	3.83
Low	2.95 (74%)	2.97 (76%)	3.47 (46%)	3.43	2.55 (100%)	2.59 (100%)	4.51 (52%)	4.52
Medium	2.94 (90%)	3.02 (88%)	3.53 (48%)	3.55	2.66 (100%)	2.78 (100%)	4.68 (40%)	4.61
High	3.01 (86%)	3.17 (76%)	3.74 (46%)	3.68	2.85 (100%)	2.83 (100%)	5.02 (28%)	4.94
Very High	2.95 (94%)	2.96 (98%)	3.63 (38%)	3.55	3.14 (100%)	3.23 (100%)	5.39 (48%)	5.39

(b) Causal Forest

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Split-only	Non-Honest	Default	Proposed	Split-only	Non-Honest	Default
No	0.71 (100%)	0.71 (100%)	1.40 (32%)	1.38	0.71 (100%)	0.71 (100%)	1.40 (32%)	1.38
Very Low	0.73 (100%)	0.74 (100%)	1.42 (33%)	1.41	1.59 (100%)	1.60 (100%)	2.85 (36%)	2.85
Low	0.78 (100%)	0.78 (100%)	1.44 (42%)	1.43	2.12 (100%)	2.14 (100%)	3.71 (48%)	3.71
Medium	0.80 (100%)	0.82 (100%)	1.47 (44%)	1.47	2.52 (100%)	2.52 (100%)	4.36 (48%)	4.36
High	0.84 (100%)	0.86 (100%)	1.47 (33%)	1.45	2.82 (100%)	2.84 (100%)	4.85 (36%)	4.84
Very High	0.84 (100%)	0.84 (100%)	1.42 (52%)	1.42	3.06 (100%)	3.09 (100%)	5.27 (32%)	5.27

Note: We calculate the RMSE from 50 random splits, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses).

Table B.35 reports the AUTOC values (multiplied by 100) for different methods across various privacy levels. Consistent with the main findings in Section 2.6 of the paper, the Proposed method outperforms all other methods, and all the post-processing methods outperform the Default approach.

Table B.35: *AUTOC Values (Multiplied by 100) for Different Methods Across Varying Privacy Levels*

(a) T-learners with Regression Forests

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Split-only	Non-Honest	Default	Proposed	Split-only	Non-Honest	Default
No	1.38 (82%)	1.36 (82%)	1.18 (60%)	1.17	1.37 (86%)	1.38 (82%)	1.18 (60%)	1.17
Very Low	1.46 (92%)	1.42 (88%)	1.17 (80%)	1.13	1.10 (76%)	1.07 (72%)	0.92 (54%)	0.91
Low	1.24 (94%)	1.23 (98%)	0.98 (82%)	0.92	1.11 (82%)	1.08 (78%)	0.90 (70%)	0.87
Medium	1.12 (90%)	1.09 (86%)	0.83 (84%)	0.76	0.99 (74%)	0.95 (68%)	0.84 (62%)	0.80
High	0.98 (92%)	0.95 (90%)	0.74 (84%)	0.69	0.97 (78%)	0.96 (66%)	0.85 (72%)	0.81
Very High	0.94 (90%)	0.91 (88%)	0.69 (84%)	0.62	0.75 (86%)	0.73 (82%)	0.56 (62%)	0.52

(b) Causal Forest

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Split-only	Non-Honest	Default	Proposed	Split-only	Non-Honest	Default
No	1.71 (60%)	1.68 (68%)	1.57 (48%)	1.59	1.71 (60%)	1.68 (68%)	1.57 (48%)	1.59
Very Low	1.55 (96%)	1.52 (84%)	1.44 (96%)	1.26	1.41 (76%)	1.43 (68%)	1.29 (60%)	1.26
Low	1.30 (100%)	1.32 (88%)	1.22 (96%)	1.03	1.56 (68%)	1.55 (72%)	1.45 (64%)	1.43
Medium	1.30 (96%)	1.21 (80%)	1.17 (96%)	0.97	1.27 (88%)	1.27 (84%)	1.18 (92%)	1.07
High	1.36 (92%)	1.29 (92%)	1.24 (98%)	0.99	1.26 (96%)	1.25 (96%)	1.02 (96%)	0.89
Very High	1.11 (100%)	1.07 (88%)	1.09 (94%)	0.89	1.06 (88%)	1.02 (78%)	1.00 (82%)	0.85

Note: We calculate the AUTOC values from 50 random splits, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses).

XGBoost-based Models.

Next, we consider two additional boosting-tree models: R-learner and T-learner with XGBoost models.

To fine-tune the hyperparameters in XGBoost models, we perform a single train-holdout split under the non-privacy setting. For conditional mean outcome models, we select parameters that minimize mean squared prediction error, while for the CATE model in the R-learner, we choose parameters that maximize the AUTOC value. The search ranges for key hyperparameters include: learning rate $\eta \in \{0.20, 0.40, 0.60, 0.80, 1.00\}$, maximum depth in each tree $\{2, 4, 6, 8, 10\}$, and the maximum number of iterations $\{10, 20, 30, 40, 50\}$. The optimal hyperparameters are as follows:

- **(R-learner)** Conditional mean outcome model: $\eta = 0.20$, maximum tree depth: 2, the maximum number of iterations: 40; CATE model: $\eta = 0.20$, maximum tree depth: 2, the maximum number of iterations: 10.

- **(T-learner)** Conditional mean outcome model for the treatment group: $\eta = 0.20$, maximum tree depth: 2, the maximum number of iterations: 20; Conditional mean outcome model for the control group: $\eta = 0.20$, maximum tree depth: 2, the maximum number of iterations: 10.

Table B.36 reports the RMSE for the two CATE models. Table B.37 reports the AUTOC values

(multiplied by 100) for different methods across various privacy levels. The findings are consistent with the main results in Section 2.6.

Table B.36: *RMSE of GATE (Multiplied by 100) for Different Methods Across Varying Privacy Levels*

(a) R-learner with XGBoost Models

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Split-only	Non-Honest	Default	Proposed	Split-only	Non-Honest	Default
No	2.16 (100%)	2.18 (100%)	2.08 (100%)	4.82	2.16 (100%)	2.18 (100%)	2.08 (100%)	4.82
Very Low	2.16 (100%)	2.18 (100%)	2.08 (100%)	5.02	2.16 (100%)	2.26 (100%)	2.34 (100%)	4.85
Low	2.16 (100%)	2.26 (100%)	2.34 (100%)	4.85	2.18 (98%)	2.39 (96%)	2.26 (100%)	4.88
Medium	2.18 (96%)	2.39 (96%)	2.26 (100%)	4.78	2.05 (100%)	2.25 (100%)	2.45 (100%)	4.85
High	2.05 (100%)	2.25 (100%)	2.45 (100%)	4.65	2.17 (100%)	2.33 (100%)	2.38 (100%)	4.88
Very High	2.17 (100%)	2.33 (100%)	2.38 (100%)	4.78	2.29 (100%)	2.32 (100%)	2.51 (100%)	4.92

(b) T-learners with XGBoost Models

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Split-only	Non-Honest	Default	Proposed	Split-only	Non-Honest	Default
No	2.48 (100%)	2.42 (100%)	2.62 (100%)	4.69	2.48 (100%)	2.42 (100%)	2.62 (100%)	4.69
Very Low	2.20 (100%)	2.23 (100%)	2.23 (100%)	4.64	2.35 (100%)	2.37 (100%)	2.47 (100%)	4.70
Low	2.11 (100%)	2.13 (100%)	2.26 (100%)	4.31	2.76 (100%)	2.57 (100%)	2.65 (100%)	4.83
Medium	2.21 (100%)	2.20 (100%)	2.24 (92%)	4.27	2.53 (100%)	2.50 (100%)	2.48 (100%)	4.65
High	2.10 (100%)	2.16 (100%)	2.16 (100%)	4.33	2.59 (100%)	2.74 (100%)	2.57 (100%)	4.60
Very High	2.21 (100%)	2.22 (100%)	2.21 (100%)	4.26	2.85 (100%)	2.96 (96%)	2.83 (96%)	4.57

Note: We calculate the RMSE from 50 random splits, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses).

Table B.37: *AUTOC Values (Multiplied by 100) for Different Methods Across Varying Privacy Levels*

(a) R-learners with XGBoost Models

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Split-only	Non-Honest	Default	Proposed	Split-only	Non-Honest	Default
No	1.78 (72%)	1.79 (84%)	1.78 (80%)	1.55	1.78 (72%)	1.79 (84%)	1.78 (80%)	1.55
Very Low	1.81 (96%)	1.83 (96%)	1.78 (96%)	1.45	1.67 (74%)	1.68 (76%)	1.61 (90%)	1.58
Low	1.71 (84%)	1.74 (88%)	1.77 (96%)	1.43	1.77 (72%)	1.75 (60%)	1.78 (80%)	1.61
Medium	1.57 (72%)	1.61 (80%)	1.57 (76%)	1.42	1.68 (68%)	1.67 (76%)	1.64 (68%)	1.55
High	1.55 (68%)	1.60 (76%)	1.62 (76%)	1.41	1.61 (64%)	1.65 (60%)	1.69 (68%)	1.56
Very High	1.54 (67%)	1.60 (75%)	1.58 (83%)	1.39	1.62 (76%)	1.67 (68%)	1.66 (76%)	1.49

(b) T-learners with XGBoost Models

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Split-only	Non-Honest	Default	Proposed	Split-only	Non-Honest	Default
No	1.68 (72%)	1.66 (70%)	1.58 (60%)	1.55	1.68 (72%)	1.66 (70%)	1.58 (60%)	1.55
Very Low	1.62 (64%)	1.68 (72%)	1.62 (64%)	1.51	1.67 (72%)	1.77 (84%)	1.57 (68%)	1.47
Low	1.74 (84%)	1.78 (96%)	1.58 (84%)	1.42	1.64 (64%)	1.60 (64%)	1.57 (72%)	1.51
Medium	1.49 (64%)	1.51 (76%)	1.47 (80%)	1.35	1.62 (76%)	1.62 (84%)	1.67 (92%)	1.46
High	1.52 (66%)	1.53 (68%)	1.56 (72%)	1.44	1.53 (84%)	1.55 (88%)	1.45 (68%)	1.31
Very High	1.52 (78%)	1.52 (80%)	1.46 (72%)	1.35	1.53 (68%)	1.54 (72%)	1.53 (72%)	1.44

Note: We calculate the AUTOC values from 50 random splits, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses).

B.7 Further Details about Starbucks Case Study

In this section, we provide the summary statistics, implementation details, and robustness checks for the Starbucks case study in Section 2.6.4.

B.7.1 Summary Statistics

Table B.38 provides summary statistics for the Starbucks data. Note that there are five categorical and two continuous pre-treatment covariates.

Table B.38: *Summary Statistics for Starbucks Data*

Discrete Variables

Variable	N	Unique Values	Distribution
Purchase (Outcome)	126,184	2	0: 124,664, 1: 1,520
Promotion (Treatment)	126,184	2	0: 63,072, 1: 63,112
V1	126,184	4	0: 15,846, 1: 47,410, 2: 47,134, 3: 15,794
V4	126,184	2	1: 40,379, 2: 85,805
V5	126,184	4	1: 23,179, 2: 46,597, 3: 48,643, 4: 7,765
V6	126,184	4	1: 31,435, 2: 31,420, 3: 3,1651, 4: 31,678
V7	126,184	2	1: 37,545, 2: 88,639

Continuous Variables

Variable	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
V2	126,184	29.98	5.00	7.10	26.596	29.98	33.354	55.108
V3	126,184	0.00	1.00	−1.69	−0.91	−0.04	0.83	1.69

B.7.2 Covariate Balance Check

Table B.39 reports the summary statistics for the covariate balance check. The result suggests that the experiment is properly randomized as the standardized mean differences of all the covariates are close to zero.

We also conduct a linear regression test to assess whether any covariate predicts the treatment assignment indicator. The joint F-test indicates that the model has no predictive power (F-stat = 0.8966, p-value = 0.5560), suggesting that randomization was successfully implemented.

B.7.3 Implementation of Local Differential Privacy

In the scenario when the outcome variable is protected by LDP, we apply the randomized response mechanism, setting p to values in the set $0.05, 0.10, 0.15, 0.20, 0.25$. In the scenario when

Table B.39: *Covariate Balance Check for Starbucks Data*

Variable	Mean Diff.	Pooled St. Dev.	Standardized Mean Diff.
V1 = 0	−0.002	0.331	−0.005
V1 = 1	−0.003	0.484	−0.006
V1 = 2	0.002	0.484	0.004
V1 = 3	0.003	0.331	0.009
V2	−0.011	5.001	−0.002
V3	0.012	1.000	0.012
V4 = 1	−0.001	0.466	−0.002
V4 = 2	0.001	0.466	0.002
V5 = 1	0.003	0.387	0.007
V5 = 2	−0.001	0.483	−0.001
V5 = 3	−0.001	0.487	−0.001
V5 = 4	−0.002	0.240	−0.006
V6 = 1	0.000	0.433	0.000
V6 = 2	0.000	0.432	0.000
V6 = 3	0.001	0.433	0.001
V6 = 4	−0.001	0.434	−0.002
V7 = 1	−0.001	0.457	−0.002
V7 = 2	0.001	0.457	0.002

covaraites are protected by LDP, we deploy the Laplace mechanism for the `V2` variable, setting $\sigma_{\texttt{V2}}$ to values in the set $5, 10, 15, 20, 25$, and for the `V3` variable, setting $\sigma_{\texttt{V3}}$ to values in the set $1, 2, 3, 4, 5$. For all discrete variables, we implement the randomized response mechanism, with p in the set $0.08, 0.16, 0.24, 0.32, 0.40$.

B.7.4 Model Specification

In the main analysis, we use the R-learner with regression forest models models and five-fold cross-fitting for the DEFAULT method and initial CATE models. We use a constant propensity score (0.5) for Robinson's transformation. For all regression forest models, we use 500 trees instead of the default 2,000 trees to accelerate training, while keeping all other parameters at their default settings in the `grf` package.

To construct the AIPW score for post-processing, we set a constant propensity score ($\widehat{e} = 0.5$), considering that the experiment is completely randomized. As for the conditional mean outcome models, we evaluate three candidate models: linear regression, logistic regression, and regression forest. The data is split into two sets: 70% allocated as the training set and the remaining 30% serving as the holdout set. The mean-squared error (MSE) metric is reported in Table B.40, calculated as $\frac{1}{N_{\text{holdout}}} \sum_{l \in \text{holdout set}} [Y_i - \widehat{m}_{W_l}(\widetilde{\mathbf{X}}_l)]^2$. Linear regression is chosen as our final model

since it yields the smallest MSE.

Table B.40: *MSE of Conditional Mean Outcome Models*

Privacy	Scenario 1: LDP-Protected Covariates			Scenario 2: LDP-Protected Outcome		
	Linear Regression	Logistic Regression	Regression Forest	Linear Regression	Logistic	Regression Forest
No	0.0119	0.0119	0.0119	0.0119	0.0119	0.0119
Very Low	0.0119	0.0119	0.0119	0.0360	0.0360	0.0360
Low	0.0114	0.0114	0.0114	0.0559	0.0559	0.0560
Medium	0.0123	0.0123	0.0123	0.0766	0.0766	0.0768
High	0.0118	0.0118	0.0118	0.0981	0.0981	0.0982
Very High	0.0117	0.0117	0.0117	0.1160	0.1160	0.1164

For the adjustment models in the post-processing procedures, we use linear regression to ensure simplicity and computational efficiency. We set the number of subgroups to 5 and the maximum number of iterations to 5 for the subgroup cross-learning algorithm.

B.7.5 Small Sample Performance

Here, we demonstrate that in small-sample settings, the Proposed method outperforms the Split-Only method. Instead of using 70% for model construction and 30% for holdout evaluation, we now only use 10% of the data to construct models and 90% to evaluate the performance.

Table B.41 presents the RMSE across varying privacy levels. The results show that the Proposed method significantly outperforms the Split-Only method, with both improving accuracy compared to the Default approach. Additionally, the Non-Honest post-processing method can even degrade performance, underscoring the importance of our Proposed method in small-sample settings.

Table B.41: *RMSE of GATE (Multiplied by 100) for Different Methods Across Varying Privacy Levels*

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Split-only	Non-Honest	Default	Proposed	Split-only	Non-Honest	Default
No	1.00 (100%)	1.03 (100%)	1.72 (20%)	1.70	1.00 (100%)	1.03 (100%)	1.72 (20%)	1.70
Very Low	0.95 (100%)	0.98 (100%)	1.67 (8%)	1.60	1.86 (100%)	1.88 (100%)	3.15 (16%)	3.12
Low	0.98 (100%)	0.99 (100%)	1.67 (16%)	1.62	2.45 (100%)	2.51 (100%)	4.12 (28%)	4.06
Medium	1.01 (100%)	1.05 (100%)	1.71 (16%)	1.64	2.94 (100%)	2.98 (100%)	4.86 (12%)	4.78
High	1.03 (100%)	1.04 (100%)	1.72 (4%)	1.65	3.29 (100%)	3.45 (100%)	5.42 (4%)	5.35
Very High	1.05 (100%)	1.06 (100%)	1.70 (4%)	1.63	3.60 (100%)	3.69 (100%)	6.00 (0%)	5.91

Note: We calculate the RMSE from 50 random splits, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses).

Table B.42 reports the AUTOC values across different privacy levels. Consistent with previous

findings, the Proposed method outperforms the Split-Only method, with both achieving better accuracy than the Default approach. While the Non-Honest post-processing method also enhances treatment prioritization ability, it suffers from higher predictive error.

Table B.42: *AUTOC (Multiplied by 100) for Different Methods Across Varying Privacy Levels*

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Split-only	Non-Honest	Default	Proposed	Split-only	Non-Honest	Default
No	0.24 (92%)	0.26 (96%)	0.24 (100%)	0.18	0.24 (92%)	0.26 (96%)	0.24 (100%)	0.18
Very Low	0.27 (92%)	0.23 (76%)	0.25 (88%)	0.18	0.14 (92%)	0.16 (88%)	0.12 (92%)	0.08
Low	0.19 (84%)	0.15 (84%)	0.15 (76%)	0.14	0.12 (72%)	0.09 (56%)	0.10 (92%)	0.07
Medium	0.15 (88%)	0.13 (68%)	0.12 (76%)	0.11	0.10 (72%)	0.09 (68%)	0.08 (84%)	0.04
High	0.12 (80%)	0.10 (68%)	0.11 (78%)	0.07	0.07 (74%)	0.03 (60%)	0.05 (84%)	0.05
Very High	0.08 (72%)	0.08 (72%)	0.07 (86%)	0.05	0.08 (84%)	0.07 (72%)	0.07 (82%)	0.04

Note: We calculate the AUTOC values from 50 random splits, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses).

B.7.6 Value Improvement for Targeting

Table B.43 reports the average value improvement across different privacy levels. The results indicate that both the Proposed method and the Split-Only method consistently outperform the default approach, with the proposed method achieving slightly better performance than the split-only method.

Table B.43: *Value Improvement Across Varying Privacy Levels: R-learners with Regression Forests*

Privacy	Scenario 1: LDP-Protected Covariates			Scenario 2: LDP-Protected Outcome		
	Proposed	Split-only	Non-Honest	Proposed	Split-only	Non-Honest
No	2.25% (88%)	2.22% (92%)	−0.03% (56%)	2.25% (88%)	2.22% (92%)	−0.03% (56%)
Very Low	4.84% (96%)	4.35% (94%)	−0.03% (48%)	2.32% (94%)	2.33% (98%)	0.06% (58%)
Low	4.8% (96%)	5.04% (96%)	−0.08% (44%)	2.48% (92%)	2.51% (94%)	0.24% (66%)
Medium	5.89% (96%)	5.48% (82%)	0.06% (56%)	2.73% (98%)	2.64% (94%)	0.21% (70%)
High	4.42% (94%)	3.94% (80%)	−0.21% (48%)	2.80% (94%)	2.85% (98%)	0.23% (72%)
Very High	5.12% (100%)	4.96% (94%)	0.37% (66%)	3.24% (100%)	3.18% (98%)	0.16% (66%)

Note: We calculate the value improvement from 50 splits, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses). The results presented here are based on the use of R-learner with regression forests as the initial CATE model.

B.7.7 Robustness Checks of Other Default Models

Forest-based Models.

We begin by considering two alternative forest-based CATE models as the DEFAULT method and the inital CATE models for the post-processing algorithms: Causal Forest and T-learner with regression forests.

1. *Causal Forest*: We use the causal forest function implemented in the `grf` package with 500 trees and other default parameters. Note that we choose 500 trees instead the default 2,000 trees to accelerate the model training process.

2. *T-learner*: We estimate two separate conditional mean outcome models — one for the treatment group and one for the control group—and compute the predicted CATEs as the difference between their predicted outcomes. For each regression forest, we use 500 trees instead of the default 2,000 trees to accelerate the training process, while keeping all other parameters at their default settings in the `grf` package.

Table B.44 reports the RMSE for the two CATE models. The results indicate that (i) the PROPOSED method significantly outperforms all other methods, (ii) the SPLIT-ONLY method also improves upon the DEFAULT method, while (iii) the NON-HONEST post-processing method fail to reduce RMSE.

Table B.45 reports the AUTOC values (multiplied by 100) for different methods across various privacy levels. Consistent with the main findings in Section 2.6 of the paper, the PROPOSED method outperforms all other methods, and all the post-processing methods outperform the DEFAULT approach.

Table B.44: *RMSE of GATE (Multiplied by 100) for Different Methods Across Varying Privacy Levels*

(a) T-learners with Regression Forests

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Split-only	Non-Honest	Default	Proposed	Split-only	Non-Honest	Default
No	0.48 (44%)	0.48 (40%)	0.46 (44%)	0.47	0.48 (44%)	0.48 (40%)	0.46 (44%)	0.47
Very Low	0.39 (56%)	0.38 (72%)	0.45 (40%)	0.44	0.43 (100%)	0.47 (100%)	0.73 (16%)	0.68
Low	0.39 (68%)	0.39 (56%)	0.45 (32%)	0.42	0.68 (100%)	0.69 (100%)	1.17 (20%)	1.14
Medium	0.40 (64%)	0.41 (52%)	0.46 (52%)	0.43	0.89 (100%)	0.95 (100%)	1.62 (20%)	1.58
High	0.38 (62%)	0.39 (72%)	0.45 (32%)	0.42	1.19 (100%)	1.23 (100%)	2.07 (40%)	2.06
Very High	0.38 (64%)	0.40 (56%)	0.44 (36%)	0.43	1.39 (100%)	1.40 (100%)	2.41 (12%)	2.39

(b) Causal Forest

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Split-only	Non-Honest	Default	Proposed	Split-only	Non-Honest	Default
No	0.46 (96%)	0.48 (88%)	0.76 (36%)	0.75	0.46 (96%)	0.48 (88%)	0.76 (36%)	0.75
Very Low	0.44 (100%)	0.44 (100%)	0.78 (24%)	0.74	0.72 (100%)	0.75 (100%)	1.37 (28%)	1.35
Low	0.42 (100%)	0.44 (100%)	0.76 (28%)	0.72	0.99 (100%)	1.00 (100%)	1.79 (52%)	1.79
Medium	0.44 (100%)	0.44 (100%)	0.78 (36%)	0.75	1.23 (100%)	1.26 (100%)	2.15 (48%)	2.14
High	0.45 (100%)	0.46 (100%)	0.78 (20%)	0.75	1.42 (100%)	1.44 (100%)	2.42 (52%)	2.41
Very High	0.44 (100%)	0.45 (100%)	0.78 (16%)	0.73	1.55 (100%)	1.57 (100%)	2.63 (48%)	2.63

Note: We calculate the RMSE from 50 random splits, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses).

Table B.45: *AUTOC Values (Multiplied by 100) for Different Methods Across Varying Privacy Levels*

(a) T-learners with Regression Forests

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Split-only	Non-Honest	Default	Proposed	Split-only	Non-Honest	Default
No	0.60 (64%)	0.59 (60%)	0.59 (68%)	0.59	0.60 (64%)	0.59 (60%)	0.59 (68%)	0.59
Very Low	0.56 (56%)	0.56 (64%)	0.56 (68%)	0.55	0.44 (76%)	0.44 (60%)	0.42 (56%)	0.42
Low	0.49 (64%)	0.48 (56%)	0.48 (64%)	0.47	0.31 (76%)	0.31 (76%)	0.30 (80%)	0.27
Medium	0.43 (64%)	0.43 (56%)	0.43 (72%)	0.42	0.27 (84%)	0.26 (68%)	0.26 (100%)	0.21
High	0.38 (80%)	0.36 (60%)	0.37 (80%)	0.36	0.18 (80%)	0.18 (76%)	0.15 (100%)	0.11
Very High	0.32 (84%)	0.31 (68%)	0.31 (80%)	0.29	0.17 (72%)	0.15 (68%)	0.14 (100%)	0.12

(b) Causal Forest

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Split-only	Non-Honest	Default	Proposed	Split-only	Non-Honest	Default
No	0.53 (72%)	0.53 (72%)	0.51 (72%)	0.50	0.53 (72%)	0.53 (72%)	0.51 (72%)	0.50
Very Low	0.51 (100%)	0.50 (100%)	0.43 (96%)	0.41	0.33 (92%)	0.33 (96%)	0.30 (96%)	0.26
Low	0.43 (100%)	0.42 (96%)	0.38 (88%)	0.35	0.20 (88%)	0.21 (86%)	0.18 (85%)	0.14
Medium	0.34 (96%)	0.33 (100%)	0.31 (92%)	0.28	0.17 (88%)	0.18 (92%)	0.16 (93%)	0.11
High	0.28 (96%)	0.28 (92%)	0.25 (96%)	0.21	0.14 (82%)	0.14 (76%)	0.12 (86%)	0.08
Very High	0.24 (90%)	0.23 (92%)	0.22 (92%)	0.18	0.14 (88%)	0.14 (80%)	0.12 (82%)	0.09

Note: We calculate the AUTOC values from 50 random splits, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses).

XGBoost-based Models.

Next, we consider two additional boosting-tree models: R-learner and T-learner with XGBoost models.

To fine-tune the hyperparameters in XGBoost models, we perform a single train-holdout split under the non-privacy setting. For conditional mean outcome models, we select parameters that minimize mean squared prediction error, while for the CATE model in the R-learner, we choose parameters that maximize the AUTOC value. The search ranges for key hyperparameters include: learning rate $\eta \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$, maximum depth in each tree $\{1, 2, 5, 10, 20\}$, and the maximum number of iterations $\{10, 20, 30, 40, 50\}$. The optimal hyperparameters are as follows:

- **(R-learner)** Conditional mean outcome model: $\eta = 0.20$, maximum tree depth: 2, the maximum number of iterations: 50; CATE model: $\eta = 0.15$, maximum tree depth: 4, the maximum number of iterations: 10.
- **(T-learner)** Conditional mean outcome model for the treatment group: $\eta = 0.15$, maximum tree depth: 2, the maximum number of iterations: 50; Conditional mean outcome model for the control group: $\eta = 0.20$, maximum tree depth: 2, the maximum number of iterations: 20.

Table B.46 reports the RMSE for the two CATE models. For R-learner, all the post-processing methods significantly outperform the Default method. For T-learner, both the Proposed and Split-Only methods outperform the Default method, while the Non-Honest post-processing method fail to reduce RMSE in the setting with LDP-protected outcomes.

Table B.47 reports the AUTOC values (multiplied by 100) for different methods across various privacy levels. For the R-learner, all post-processing methods significantly outperform the Default method. In contrast, for the T-learner, all methods yield similarly high AUTOC values (relative to other CATE models). This is likely because the best cross-validated XGBoost models in this setting are already simple and less prone to high variance, leaving limited room for further improvement in treatment prioritization.

Table B.46: *RMSE of GATE (Multiplied by 100) for Different Methods Across Varying Privacy Levels*

(a) R-learner with XGBoost Models

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Split-only	Non-Honest	Default	Proposed	Split-only	Non-Honest	Default
No	0.46 (100%)	0.48 (100%)	0.48 (100%)	9.62	0.46 (100%)	0.48 (100%)	0.48 (100%)	9.62
Very Low	0.50 (100%)	0.47 (100%)	0.45 (100%)	9.65	0.53 (100%)	0.59 (100%)	0.52 (100%)	9.63
Low	0.48 (100%)	0.43 (100%)	0.43 (100%)	9.65	0.58 (100%)	0.71 (100%)	0.58 (100%)	9.69
Medium	0.48 (100%)	0.44 (100%)	0.45 (100%)	9.66	0.66 (100%)	0.8 (100%)	0.70 (100%)	9.61
High	0.47 (100%)	0.41 (100%)	0.41 (100%)	9.61	0.73 (100%)	0.9 (100%)	0.77 (100%)	9.56
Very High	0.49 (100%)	0.42 (100%)	0.41 (100%)	9.66	0.84 (100%)	1.01 (100%)	0.92 (100%)	9.52

(b) T-learner with XGBoost Models

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Split-only	Non-Honest	Default	Proposed	Split-only	Non-Honest	Default
No	0.48 (94%)	0.45 (94%)	0.98 (10%)	0.68	0.48 (94%)	0.45 (94%)	0.98 (10%)	0.68
Very Low	0.44 (100%)	0.46 (96%)	0.43 (92%)	0.69	0.58 (100%)	0.60 (86%)	1.00 (36%)	0.78
Low	0.45 (96%)	0.48 (96%)	0.43 (100%)	0.69	0.69 (86%)	0.73 (72%)	1.12 (28%)	0.79
Medium	0.44 (92%)	0.45 (92%)	0.43 (96%)	0.70	0.81 (78%)	0.89 (46%)	1.22 (18%)	0.88
High	0.44 (96%)	0.44 (96%)	0.44 (96%)	0.64	0.84 (64%)	0.91 (38%)	1.27 (12%)	0.87
Very High	0.45 (100%)	0.48 (96%)	0.43 (100%)	0.70	0.94 (66%)	0.99 (42%)	1.32 (14%)	0.97

Note: We calculate the RMSE from 50 random splits, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses).

Table B.47: *AUTOC Values (Multiplied by 100) for Different Methods Across Varying Privacy Levels*

(a) R-learners with XGBoost Models

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Split-only	Non-Honest	Default	Proposed	Split-only	Non-Honest	Default
No	0.62 (64%)	0.60 (44%)	0.62 (56%)	0.61	0.62 (64%)	0.60 (44%)	0.62 (56%)	0.61
Very Low	0.60 (65%)	0.56 (60%)	0.55 (80%)	0.55	0.57 (56%)	0.52 (38%)	0.57(56%)	0.56
Low	0.48 (72%)	0.49 (76%)	0.50 (84%)	0.47	0.47 (80%)	0.41 (36%)	0.39 (28%)	0.44
Medium	0.42 (80%)	0.42 (72%)	0.42 (80%)	0.42	0.39 (84%)	0.38 (64%)	0.38 (68%)	0.33
High	0.39 (74%)	0.39 (76%)	0.39 (72%)	0.36	0.35 (82%)	0.30 (64%)	0.30 (56%)	0.26
Very High	0.31 (76%)	0.31 (68%)	0.30 (70%)	0.29	0.33 (76%)	0.29 (68%)	0.32 (74%)	0.25

(b) T-learners with XGBoost Models

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Split-only	Non-Honest	Default	Proposed	Split-only	Non-Honest	Default
No	0.67 (54%)	0.66 (42%)	0.67 (60%)	0.66	0.67 (54%)	0.66 (42%)	0.67 (60%)	0.66
Very Low	0.55 (68%)	0.56 (68%)	0.55 (76%)	0.54	0.61 (48%)	0.60 (38%)	0.63 (74%)	0.61
Low	0.50 (66%)	0.49 (44%)	0.51 (66%)	0.49	0.51 (46%)	0.50 (40%)	0.53 (62%)	0.52
Medium	0.44 (72%)	0.43 (56%)	0.44 (80%)	0.42	0.48 (52%)	0.46 (48%)	0.51 (78%)	0.47
High	0.40 (64%)	0.38 (36%)	0.40 (66%)	0.38	0.41 (46%)	0.40 (42%)	0.44 (82%)	0.42
Very High	0.30 (50%)	0.30 (48%)	0.30 (52%)	0.30	0.39 (52%)	0.37 (40%)	0.44 (82%)	0.38

Note: We calculate the AUTOC values from 50 random splits, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses).

Appendix C

Appendix to Chapter 3

C.1 Theoretical Guarantees

In this appendix, we provide the theoretical guarantees for the proposed deep IRL framework.

C.1.1 Formalizing the IRL Problem

We begin by formalizing the incrementality representation learning problem. First, we assume that the true CATE function can be decomposed into two distinct functions: the representation function, which transforms raw design features and customer covariates into a representation, and the prediction function, which maps this representation to the actual CATE. This assumption is articulated as follows:

Assumption C.1 (Data Generating Process) Assume that the true CATE function defined in Equation (3.1) can be written as

$$\tau(\mathbf{Z}_o, \mathbf{X}_{o,i}) = f^\star \circ \mathbf{h}^\star(\mathbf{Z}_o, \mathbf{X}_{o,i}),$$

where $\mathbf{h}^\star : \mathcal{Z} \times \mathcal{X} \to \mathbb{R}^K$ denotes the representation function and $f^\star : \mathbb{R}^K \to [-r, r], r < \infty$ is the prediction function for each offer. Also, assume that the probability of the firm choosing a specific design feature $\mathbf{Z} \in \mathcal{Z}$ and observing specific customer covariate values $\mathbf{X} \subset \mathcal{X}$ are non-zero for an experiment.

Note that this assumption does not require the existence of more predictive low-dimensional

representations. Specifically, if the original design features and customer covariates already constitute the most meaningful representation, then we have $\mathbf{h}^\star(\mathbf{Z}_o, \mathbf{X}_{o,i}) = (\mathbf{Z}_o, \mathbf{X}_{o,i})$.

Assume that we have access to an unbiased proxy $\widetilde{\tau}_{o,i}$ for the actual CATE (e.g., estimated through the doubly robust score in Theorem 3.1). Without loss of generality and for simplicity, we assume that each past experiment has the same sample size N. The objective of IRL is to construct $\widehat{\mathbf{h}}$ and $\widehat{f}$ that best recover $\mathbf{h}^\star$ and $f^\star$ by learning from $\widetilde{\tau}_{o,i}$, that is,

$$(\widehat{f}, \widehat{\mathbf{h}}) = \underset{f \in \mathcal{F},\ \mathbf{h} \in \mathcal{H}}{\arg\min} \ \frac{1}{ON} \sum_{o=1}^{O} \sum_{i \in \mathcal{C}_o^E} \ell \left(f \circ \mathbf{h}(\mathbf{Z}_o, \mathbf{X}_{o,i}), \widetilde{\tau}_{o,i} \right). \tag{C.1}$$

Here, ℓ represents a loss function, while $\mathcal{H}$ and $\mathcal{F}$ denote the classes for the representation and prediction models, respectively (such as deep neural networks). In the following analysis, we focus on the case where $\ell(y, y') = (y - y')^2$ is the squared-loss function.

C.1.2 Model Complexity

Before establishing the bound on the prediction error, we first introduce a widely used metric to assess the complexity of a function class in machine learning theory, known as *Gaussian complexity* [22]. This measure quantifies the capacity of a model class to learn various patterns from data.

Definition C.1 (Gaussian Complexity) For a generic function class $\mathcal{Q}$, comprising functions $\mathbf{q}(\cdot) = (q_1(\cdot), \cdots, q_B(\cdot)) : \mathbb{R}^A \to \mathbb{R}^B$, and given N observed data points, $\bar{\mathbf{x}} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)$, the *empirical Gaussian complexity* of the function class is defined as

$$\widehat{\mathfrak{G}}_{\bar{\mathbf{x}}}(\mathcal{Q}) = \mathbb{E}_{\varepsilon_{n,b}} \left[\sup_{\mathbf{q} \in \mathcal{Q}} \frac{1}{N} \sum_{n=1}^{N} \sum_{b=1}^{B} \varepsilon_{n,b}\, q_b(\mathbf{x}_n) \right], \quad \varepsilon_{n,b} \overset{\text{iid}}{\sim} \mathcal{N}(0,1).$$

Following this, the population Gaussian complexity is defined as $\mathfrak{G}_N(\mathcal{Q}) = \mathbb{E}_{\bar{\mathbf{x}}} \left[\widehat{\mathfrak{G}}_{\bar{\mathbf{x}}}(\mathcal{Q}) \right]$, and the worst-case emprical Gaussian complexity is defined as $\overline{\mathfrak{G}}_N(\mathcal{Q}) = \arg\max_{\bar{\mathbf{x}}} \widehat{\mathfrak{G}}_{\bar{\mathbf{x}}}(\mathcal{Q})$.

Essentially, Gaussian complexity measures how closely a function class $\mathcal{Q}$ can correlate with white noise. A higher Gaussian complexity signifies a greater ability of the function class to fit random noise, indicating an increased risk of overfitting.

C.1.3 Bounding the Prediction Error of IRL

We now provide the prediction error bound for the IRL problem and its specific application to IRL models utilizing deep neural networks.

Prediction Error Bound for Generic IRL Model. We characterize the prediction error as the average mean squared error between the predictions of an IRL model and the true CATE:

$$\text{Error}\left(f, \mathbf{h}\right) = \mathbb{E}\left[\frac{1}{O}\sum_{o=1}^{O} \ell\left(f \circ \mathbf{h}(\mathbf{Z}_o, \mathbf{X}_{o,i}), f^\star \circ \mathbf{h}^\star(\mathbf{Z}_o, \mathbf{X}_{o,i})\right)\right].$$

Throughout the analysis, we impose the following standard regularity assumptions on the loss function and CATE functions.

Assumption C.2 (Assumption 1 in [193])

1. *The loss function ℓ is nonnegative and B-bounded, i.e., $0 \leqslant \ell(y', y) \leqslant B$ for any $y, y' \in \mathcal{R}$.*
2. *Every function within $\mathcal{F}$ is L-Lipschitz w.r.t. the norm $\|\cdot\|_2$, i.e., $\|f(\mathbf{a}) - f(\mathbf{a}')\|_2 \leqslant L\|\mathbf{a} - \mathbf{a}'\|_2$ for all $f \in \mathcal{F}$ and $\mathbf{a}, \mathbf{a}' \in dom(f)$.*
3. *The composed function $f \circ \mathbf{h}$ is D-bounded, i.e., $\sup_{\mathbf{Z},\mathbf{X}} |f \circ \mathbf{h}(\mathbf{Z}, \mathbf{X})| \leqslant D$.*

We now present the upper bound for the prediction error for a generic IRL model.

Theorem C.1 (Upper Bound for Prediction Error) *Assuming that (i) the regularity conditions in Assumption C.2 hold, and (ii) $f^\star \in \mathcal{F}$ and $\mathbf{h}^\star \in \mathcal{H}$[1], then with probability at least $1 - \delta$, the prediction error of the IRL model $(\widehat{f}, \widehat{\mathbf{h}})$ derived from* (C.1) *can be bounded as follows:*

$$\text{Error}\left(\widehat{f}, \widehat{\mathbf{h}}\right) \leqslant \underbrace{\frac{C_1}{(ON)^2} + \log(ON)\left[C_2 \mathfrak{G}_{ON}(\mathcal{H}) + \frac{1}{\sqrt{O}}\overline{\mathfrak{G}}_{ON}(\mathcal{F})\right]}_{\textit{Overfitting Potential for the Model Classes}} + C_3\sqrt{\frac{\log(2/\delta)}{ON}}, \tag{C.2}$$

where C_1, C_2, and C_3 are some constants.

Proof. See C.2.3. ■

[1]Although this assumption might appear strong at first glance, it is actually reasonable for deep neural networks. Previous research has demonstrated the universal approximation property of neural networks [e.g., 50, 107, 162, 139, 129], highlighting their capability to approximate a wide variety of function classes effectively.

Theorem C.1 indicates that the prediction error of an IRL model is limited by the Gaussian complexity related to (i) the class of potential representation functions ($\mathcal{H}$) and (ii) the class of potential prediction functions ($\mathcal{F}$). Therefore, for the IRL model to accurately replicate the true functions $(f^\star, \mathbf{h}^\star)$ with large O (number of past experiments) and N (sample size of each experiment), it is crucial that the Gaussian complexity of both the representation and prediction model classes declines to zero at a sufficiently fast rate.

Prediction Error Bound for DNN-based IRL Model. Next, we present the prediction error bound for the model classes constituted by depth-d vector-valued neural networks. Specifically, a neural network is formulated as:

$$\mathbf{NN}(\mathbf{x}) = \mathbf{W}_d \, \sigma_d \left(\mathbf{W}_{d-1} \, \sigma_{d-1} \left(\cdots \mathbf{W}_1 \, \sigma_1 \left(\mathbf{x} \right) \right) \right),$$

where $\sigma_1, \cdots, \sigma_d$ represent 1-Lipschitz activation functions that are zero at the origin, such as the linear and ReLU functions. Here, we define $M(j)$ as the maximum possible value of the Frobenius norm for the matrix $\mathbf{W}_j$, that is, the square root of the sum of the squares of all elements in $\mathbf{W}_j$.

Corollary C.1 (Prediction Error Bound of Deep IRL) *Consider the case when $\mathcal{H}$ is the class of d_1-layer neural networks with 1-Lipschitz activation functions and K'-dimensional outputs, and $\mathcal{F}$ denote a class of neural networks with d_2 layers, also employing 1-Lipschitz activation functions. Furthermore, it is assumed that $M(j) < \infty$ for all parameter matrices. Then, with probability at least $1 - \delta$, the prediction error of the IRL model $(\widehat{f}, \widehat{\mathbf{h}})$ derived from* (C.1) *can be bounded as follows:*

$$\begin{aligned} \text{Error}\left(\widehat{f}, \widehat{\mathbf{h}}\right) \leqslant \frac{C_1}{(ON)^2} + C_2\sqrt{\frac{\log(2/\delta)}{ON}} + \\ \log(ON)\left[C_3 \frac{\sqrt{\log(ON)} K' \prod_{j=1}^{d_1} M(j)}{\sqrt{ON}} + C_4(K') \frac{\sqrt{\log(ON)} \prod_{j=1}^{d_2} M(j)}{\sqrt{O^2 N}}\right], \end{aligned}$$

where C_1, C_2, C_3 and $C_4(K')$ are some constants. Note that $C_4(K') = \mathcal{O}(\sqrt{K'})$.

Proof. See C.2.4. ■

Corollary C.1 elucidates several critical insights regarding the application of deep neural networks within the IRL framework. Firstly, it establishes a theoretical guarantee, indicating that the upper bound of the prediction error for a deep learning-based IRL model diminishes to zero as both O

(the number of offers) and N (the sample size) increase. Secondly, it highlights the advantage of dimension reduction in deep representation learning. The prediction error bound improves with a reduction in K', provided the neural network of K' dimensions can accurately capture the true representation function $\mathbf{h}^\star$.

Thirdly, it demonstrates the benefits of utilizing separate network structures for offer and customer representations. Specifically, such neural network architectures often result in many zero values in the parameter matrix $\mathbf{W}_j$ that links the nodes between the offer representation and customer representation networks. In scenarios with separate network structures, the maximum possible value of the Frobenius norm of $\mathbf{W}_j$ is generally lower compared to a fully connected network that permits interactions between design features and customer covariates, assuming the same number of nodes. Consequently, the prediction error bound is tighter for models with separate network structures as opposed to fully connected configurations, maintaining equal depth and node count in each layer. This rationale extends to the comparison between shared-parameter prediction networks and offer-specific prediction networks, with the former typically achieving more favorable error bounds under similar conditions.

C.2 Proofs

C.2.1 Proof for Identification Results in Equation (3.2)

By the unconfoundedness and no interference assumption, we have

$$\mathbb{E}_{i\in\mathcal{C}_o^E}[Y_{o,i}|W_{o,i}=1,\mathbf{X}_{o,i}]=\mathbb{E}_{i\in\mathcal{C}_o^E}[Y_{o,i}(\mathbf{Z}_o)|\mathbf{X}_{o,i}],$$
$$\mathbb{E}_{i\in\mathcal{C}_o^E}[Y_{o,i}|W_{o,i}=0,\mathbf{X}_{o,i}]=\mathbb{E}_{i\in\mathcal{C}_o^E}[Y_{o,i}(\mathbf{0})|\mathbf{X}_{o,i}].$$

The above equations hold for all the customers in the target population $\mathcal{C}_o^E$ by the overlap assumption.

Then, by the stability assumption, we have

$$\mathbb{E}_{i\in\mathcal{C}_o^E}[Y_{o,i}(\mathbf{Z}_o)|\mathbf{X}_{o,i}]=\mathbb{E}_{i\in\mathcal{C}_o}[Y_{o,i}(\mathbf{Z}_o)|\mathbf{X}_{o,i}],$$
$$\mathbb{E}_{i\in\mathcal{C}_o^E}[Y_{o,i}(\mathbf{0})|\mathbf{X}_{o,i}]=\mathbb{E}_{i\in\mathcal{C}_o}[Y_{o,i}(\mathbf{0})|\mathbf{X}_{o,i}].$$

Combining these two results, we prove the identification results in Equation (3.2).

C.2.2 Proof for Theorem 3.1

Under Assumption 3.1, we have $\mu_{o,1}(\mathbf{X}_{o,i}) = \mathbb{E}_{\mathcal{C}_o}[Y_{o,i}(\mathbf{Z}_o)|\mathbf{X}_{o,i}]$ and $\mu_{o,0}(\mathbf{X}_{o,i}) = \mathbb{E}_{\mathcal{C}_o}[Y_{o,i}(\mathbf{0})|\mathbf{X}_{o,i}]$. Therefore, we have

$$\tau(\mathbf{Z}_o, \mathbf{X}_{o,i}) = \mu_{o,1}(\mathbf{X}_{o,i}) - \mu_{o,0}(\mathbf{X}_{o,i}).$$

Next, by definition of the propensity score $\pi_o(\mathbf{X}_{o,i})$, we have

$$\mathbb{E}_{\mathcal{C}_o}\left[W_i|\mathbf{X}_{o,i}, \mathbf{Z}_o\right] = \mathbb{P}_{\mathcal{C}_o}\left[W_i = 1|\mathbf{X}_{o,i}, \mathbf{Z}_o\right] \cdot 1 + 0 = \pi_o(\mathbf{X}_{o,i}).$$

Under Assumption 3.1, we have

$$\mathbb{E}_{\mathcal{C}_o^E}\left[Y_{o,i}|\mathbf{X}_{o,i}, \mathbf{Z}_o\right] = \mu_{o,W_{o,i}}(\mathbf{X}_{o,i}).$$

These two results implies that

$$\mathbb{E}\left[\frac{W_{o,i} - \pi_o(\mathbf{X}_{o,i})}{\pi_o(\mathbf{X}_{o,i})[1 - \pi_o(\mathbf{X}_{o,i})]}\left[Y_{o,i} - \mu_{o,W_{o,i}}(\mathbf{X}_{o,i})\right]\middle|\mathbf{X}_{o,i}, \mathbf{Z}_o\right] = 0.$$

Finally, combining these results proves Theorem 3.1.

C.2.3 Proof for Theorem C.1

We now prove Theorem C.1, which follows closely the proof for Theorem 1 in [193]. For a given f and $\mathbf{h}$, we write the population loss as

$$l(f, \mathbf{h}, f^\star, \mathbf{h}^\star) = \frac{1}{O}\,\mathbb{E}\left[\sum_{o=1}^{O} \ell\left(f \circ \mathbf{h}(\mathbf{Z}_o, \mathbf{X}_{o,i}), \widetilde{\tau}_{o,i}\right) - \ell\left(f^\star \circ \mathbf{h}^\star(\mathbf{Z}_o, \mathbf{X}_{o,i}), \widetilde{\tau}_{o,i}\right)\right],$$

where the expectation is taken over the data collection mechanism for past experiments.

Claim. $l(f, \mathbf{h}, f^\star, \mathbf{h}^\star) = \frac{1}{O}\,\mathbb{E}\left[\sum_{o=1}^{O} \ell\left(f \circ \mathbf{h}(\mathbf{Z}_o, \mathbf{X}_{o,i}), f^\star \circ \mathbf{h}^\star(\mathbf{Z}_o, \mathbf{X}_{o,i})\right)\right]$.

Proof. We first note that

$$\begin{aligned}
&\mathbb{E}\left[\ell\left(f \circ \mathbf{h}(\mathbf{Z}_o, \mathbf{X}_{o,i}), \widetilde{\tau}_{o,i}\right)|\mathbf{Z}_o, \mathbf{X}_{o,i}\right] \\
&= \left[f \circ \mathbf{h}(\mathbf{Z}_o, \mathbf{X}_{o,i})\right]^2 - 2f \circ \mathbf{h}(\mathbf{Z}_o, \mathbf{X}_{o,i})\mathbb{E}\left[\widetilde{\tau}_{o,i}|\mathbf{Z}_o, \mathbf{X}_{o,i}\right] + \mathbb{E}^2\left[\widetilde{\tau}_{o,i}|\mathbf{Z}_o, \mathbf{X}_{o,i}\right] + \mathrm{Var}\left[\widetilde{\tau}_{o,i}|\mathbf{Z}_o, \mathbf{X}_{o,i}\right] \\
&= \left[f \circ \mathbf{h}(\mathbf{Z}_o, \mathbf{X}_{o,i})\right]^2 - 2f \circ \mathbf{h}(\mathbf{Z}_o, \mathbf{X}_{o,i}) \cdot f^\star \circ \mathbf{h}^\star(\mathbf{Z}_o, \mathbf{X}_{o,i}) + \left[f^\star \circ \mathbf{h}^\star(\mathbf{Z}_o, \mathbf{X}_{o,i})\right]^2 + \mathrm{Var}\left[\widetilde{\tau}_{o,i}|\mathbf{Z}_o, \mathbf{X}_{o,i}\right].
\end{aligned}$$

for given f and $\mathbf{h}$ since $\widetilde{\tau}_{o,i}$ is unbiased. Similarly, we can show that

$$\mathbb{E}\left[\ell\left(f^{\star}\circ\mathbf{h}^{\star}(\mathbf{Z}_o,\mathbf{X}_{o,i}),\widetilde{\tau}_{o,i}\right)|\mathbf{Z}_o,\mathbf{X}_{o,i}\right]=\mathrm{Var}\left[\widetilde{\tau}_{o,i}|\mathbf{Z}_o,\mathbf{X}_{o,i}\right].$$

Combining these two observations, we have

$$\begin{aligned}
&\mathbb{E}\left[\ell\left(f\circ\mathbf{h}(\mathbf{Z}_o,\mathbf{X}_{o,i}),\widetilde{\tau}_{o,i}\right)-\ell\left(f^{\star}\circ\mathbf{h}^{\star}(\mathbf{Z}_o,\mathbf{X}_{o,i}),\widetilde{\tau}_{o,i}\right)|\mathbf{Z}_o,\mathbf{X}_{o,i}\right]\\
&\quad=\left[f\circ\mathbf{h}(\mathbf{Z}_o,\mathbf{X}_{o,i})\right]^2-2f\circ\mathbf{h}(\mathbf{Z}_o,\mathbf{X}_{o,i})\cdot f^{\star}\circ\mathbf{h}^{\star}(\mathbf{Z}_o,\mathbf{X}_{o,i})+\left[f^{\star}\circ\mathbf{h}^{\star}(\mathbf{Z}_o,\mathbf{X}_{o,i})\right]^2\\
&\quad=\ell\left(f\circ\mathbf{h}(\mathbf{Z}_o,\mathbf{X}_{o,i}),f^{\star}\circ\mathbf{h}^{\star}(\mathbf{Z}_o,\mathbf{X}_{o,i})\right).
\end{aligned}$$

By taking expectation of the above results for $\mathbf{Z}_o,\mathbf{X}_{o,i}$, we can prove the claim. ■

Next, we can define the empirical loss as:

$$\widehat{l}(f,\mathbf{h},f^{\star},\mathbf{h}^{\star})=\frac{1}{ON}\sum_{o=1}^{O}\sum_{i\in\mathcal{C}_o^E}\ell\left(f\circ\mathbf{h}(\mathbf{Z}_o,\mathbf{X}_{o,i}),\widetilde{\tau}_{o,i}\right)-\mathbb{E}\left[\ell\left(f^{\star}\circ\mathbf{h}^{\star}(\mathbf{Z}_o,\mathbf{X}_{o,i}),\widetilde{\tau}_{o,i}\right)\right].$$

Now, let $\widehat{f}$ and $\widehat{\mathbf{h}}$ be derived from the empirical risk minimization problem in (C.1) of the main document. Then, we have

$$\begin{aligned}
l(\widehat{f},\widehat{\mathbf{h}},f^{\star},\mathbf{h}^{\star})&=l(\widehat{f},\widehat{\mathbf{h}},f^{\star},\mathbf{h}^{\star})-\underbrace{l(f^{\star},\mathbf{h}^{\star},f^{\star},\mathbf{h}^{\star})}_{=0}\\
&=\underbrace{l(\widehat{f},\widehat{\mathbf{h}},f^{\star},\mathbf{h}^{\star})-\widehat{l}(\widehat{f},\widehat{\mathbf{h}},f^{\star},\mathbf{h}^{\star})}_{\equiv a}+\underbrace{\widehat{l}(\widehat{f},\widehat{\mathbf{h}},f^{\star},\mathbf{h}^{\star})-\widehat{l}(f^{\star},\mathbf{h}^{\star},f^{\star},\mathbf{h}^{\star})}_{\equiv b}+\\
&\quad\underbrace{\widehat{l}(f^{\star},\mathbf{h}^{\star},f^{\star},\mathbf{h}^{\star})-l(f^{\star},\mathbf{h}^{\star},f^{\star},\mathbf{h}^{\star})}_{\equiv c}\leqslant a+c.
\end{aligned}$$

Note that the inequality holds as $b<0$ by the fact that $\widehat{f}$ and $\widehat{\mathbf{h}}$ are derived by minimizing $\widehat{l}(\widehat{f},\widehat{\mathbf{h}},f^{\star},\mathbf{h}^{\star})$.

We can now establish bounds for the population prediction error of the estimator. Using the standard generalization bounds based on Rademacher complexity [22, 200], we can bound the two components a and c as:

$$a,c\leqslant\sup_{f\in\mathcal{F},\mathbf{h}\in\mathcal{H}}\left|l(f,h,f^{\star},\mathbf{h}^{\star})-\widehat{l}(f,h,f^{\star},\mathbf{h}^{\star})\right|\leqslant 2\mathfrak{R}_{ON}(\ell(\mathcal{F}(\mathcal{H})))+2B\sqrt{\frac{\log(2/\delta)}{ON}}$$

with probability at $1-\delta$, where $\mathcal{F}(\mathcal{H})=\{(f,\mathbf{h}):f\in\mathcal{F}\text{ and }\mathbf{h}\in\mathcal{H}\}$, and $\mathfrak{R}_{ON}(\ell(\mathcal{F}(\mathcal{H})))$ is the

Rademacher complexity of $\ell(\mathcal{F}(\mathcal{H}))$. The first inequality is a direct consequence of the definition of sup. The second inequality leverages the well-established probabilistic upper bound derived using the Rademacher complexity [200].

Then, by Inequality (11) in the Appendix of [193], we have

$$\mathfrak{R}_{ON}(\ell(\mathcal{F}(\mathcal{H}))) \leqslant 2L\mathfrak{R}_{ON}(\mathcal{F}(\mathcal{H})).$$

Note that $\mathfrak{R}_{ON}(\mathcal{F}(\mathcal{H})) \leqslant \sqrt{\frac{\pi}{2}}\mathfrak{G}_{ON}(\mathcal{F}(\mathcal{H}))$. Combining the above results together, we can establish the upper bound for the population prediction error:

$$\begin{aligned} l(\widehat{f}, \widehat{\mathbf{h}}, f^\star, \mathbf{h}^\star) \leqslant a + c &\leqslant 4\mathfrak{R}_{ON}(\ell(\mathcal{F}(\mathcal{H}))) + 4B\sqrt{\frac{\log(2/\delta)}{ON}} \\ &\leqslant 8L\mathfrak{R}_{ON}(\mathcal{F}(\mathcal{H})) + 8B\sqrt{\frac{\log(2/\delta)}{ON}} \\ &\leqslant 16L\mathfrak{G}_{ON}(\mathcal{F}(\mathcal{H})) + 8B\sqrt{\frac{\log(2/\delta)}{ON}}. \end{aligned} \tag{C.3}$$

The last inequality holds since the Rademacher complexity is upper bounded by the Gaussian complexity [135]: $\mathfrak{R}_{ON}(\mathcal{F}(\mathcal{H})) \leqslant \sqrt{\frac{\pi}{2}}\mathfrak{G}_{ON}(\mathcal{F}(\mathcal{H})) < 2\mathfrak{G}_{ON}(\mathcal{F}(\mathcal{H}))$.

Now, we can further decompose the Gaussian complexity based on Theorem 7 in [193]. Note that the Gaussian complexity is $\mathfrak{G}_{ON}(\mathcal{F}(\mathcal{H}))$ in our setting instead of $\mathfrak{G}_{ON}(\mathcal{F}^{\otimes O}(\mathcal{H}))$ in [193]. Therefore, when bounding the covering number for the Dudley entropy integral bound (page 18 of [193]), we calculate the cadinality for the covering $\mathcal{C}_{\mathcal{F}(\mathcal{H})}$ of the set $\mathcal{F}(\mathcal{H})$, rather than the covering $\mathcal{C}_{\mathcal{F}^{\otimes O}(\mathcal{H})}$ of the set $\mathcal{F}^{\otimes O}(\mathcal{H})$. Therefore, the last inequality in page 18 of [193] becomes:

$$\log N_{2,\mathbf{X}}\left(\varepsilon_1 \cdot \mathcal{L}(\mathcal{F}) + \varepsilon_2, d_{2,\mathbf{X}}, \mathcal{F}(\mathcal{H})\right) \leqslant \log N_{2,\mathbf{X}}\left(\varepsilon_1, d_{2,\mathbf{X}}, \mathcal{H}\right) + \max_{\mathbf{z}\in\mathcal{Z}} \log N_{2,\mathbf{z}}\left(\varepsilon_2, d_{2,\mathbf{z}}, \mathcal{F}\right).$$

Applying this result to the rest of proof for Theorem 7 in [193] gives the following inequality:

$$\mathfrak{G}_{ON}(\mathcal{F}(\mathcal{H})) \leqslant 64\frac{D}{(ON)^2} + 128C(\mathcal{F}(\mathcal{H}))\log(ON), \tag{C.4}$$

where $C(\mathcal{F}(\mathcal{H})) = L\mathfrak{G}_{ON}(\mathcal{H}) + \frac{1}{\sqrt{O}}\max_{\mathbf{q}\in\{\mathbf{h}(\mathbf{Z}_o,\mathbf{X}_{o,i}):\mathbf{h}\in\mathcal{H}}\ \widehat{\mathfrak{G}}_{\mathbf{q}}(\mathcal{F})$.

Finally, combining (C.4) and (C.3) gives

$$\text{Error}\left(\widehat{f}, \widehat{\mathbf{h}}\right) \leqslant \underbrace{\frac{C_1}{(ON)^2} + \log(ON)\left[C_2\mathfrak{G}_{ON}(\mathcal{H}) + \frac{1}{\sqrt{O}}\overline{\mathfrak{G}}_{ON}(\mathcal{F})\right]}_{\text{Overfitting Potential for the Model Classes}} + C_3\sqrt{\frac{\log(2/\delta)}{ON}},$$

with probability at least $(1-\delta)$.

C.2.4 Proof for Corollary C.1

By Theorem 2 in [85], we can bound the sample Rademacher complexity of deep neural networks $\mathcal{NN}$ with K-dimensional output and depth d as follows:

$$\widehat{\mathfrak{R}}_{ON}(\mathcal{NN}) \leqslant \frac{2K\sqrt{d+1+\log(r)}\prod_{j=1}^{j} M(j)}{\sqrt{ON}}, \tag{C.5}$$

where $M(j)$ is the maximum possible value of the Frobenius norm for the weight matrix in the j-th layer, and r is the dimension of the input variables.

Using the fact that $\widehat{\mathfrak{G}}_{ON}(\mathcal{NN}) \leqslant 2\sqrt{\log(ON)}\,\widehat{\mathfrak{R}}_{ON}(\mathcal{NN})$ (page 97 in [135]) together with (C.5) and taking expectation on the left side, we have

$$\mathfrak{G}(\mathcal{NN})_{ON} \leqslant \frac{4\sqrt{\log(ON)}K\sqrt{d+1+\log(r)}\prod_{j=1}^{j} M(j)}{\sqrt{ON}}.$$

As a result, for the function classes $\mathcal{H}$ and $\mathcal{F}$ specified in the corollary, we have the following bounds for their Gaussian complexities:

$$\mathfrak{G}_{ON}(\mathcal{H}) \leqslant A\frac{\sqrt{\log(ON)}K'\prod_{j=1}^{j} M(j)}{\sqrt{ON}}, \overline{\mathfrak{G}}_{ON}(\mathcal{F}) \leqslant B\sqrt{d+1+\log(K')}\frac{\sqrt{\log(ON)}\prod_{j=1}^{j} M(j)}{\sqrt{ON}}.$$

Finally, by applying the above bounds of Gaussian complexities on Theorem C.1, we can prove the corollary.

C.2.5 Proof for Theorem 3.2

To prove the theorem, we first note that

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_{\text{new}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{new}}\right)\right] &= \mathbb{E}_{\mathcal{D}_{\text{new}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{new}}\right)\right] + \\
&\quad \mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{past}}\right)\right] - \mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{past}}\right)\right] + \\
&\quad \mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{new}}\right)\right] - \mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{new}}\right)\right] \\
&\leqslant \mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{past}}\right)\right] + \\
&\quad \underbrace{\left|\mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{new}}\right)\right] - \mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{past}}\right)\right]\right|}_{=a} + \\
&\quad \underbrace{\left|\mathbb{E}_{\mathcal{D}_{\text{new}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{new}}\right)\right] - \mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{past}}\right)\right]\right|}_{=b}
\end{aligned}$$

For the first term, by the reverse triangle inequality — $|d(x,y) - d(x,z)| \leqslant d(y,x)$ for any valid distance metric d — we have $a \leqslant \mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\tau_{\text{past}}, \tau_{\text{new}}\right)\right]$.

For the second term, we have

$$b \leqslant \int \left|\rho_{\mathcal{D}_{\text{new}}}(\mathbf{z}, \mathbf{x}) - \rho_{\mathcal{D}_{\text{past}}}(\mathbf{z}, \mathbf{x})\right| \ell\left(\widehat{\tau}, \tau_{\text{new}}\right) \leqslant C\delta(\mathcal{D}_{\text{past}}, \mathcal{D}_{\text{new}}),$$

where $\rho_{\mathcal{D}_{\text{new}}}$ and $\rho_{\mathcal{D}_{\text{past}}}$ are the density functions of the two data generating processes. Note that the term $\ell\left(\widehat{\tau}, \tau_{\text{new}}\right)$ can be bounded by C as $\widehat{\tau}, \tau_{\text{new}}$ are boubded by assumption. Therefore, we have

$$\mathbb{E}_{\mathcal{D}_{\text{new}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{new}}\right)\right] \leqslant \mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{past}}\right)\right] + \mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\tau_{\text{past}}, \tau_{\text{new}}\right)\right] + C\delta(\mathcal{D}_{\text{past}}, \mathcal{D}_{\text{new}}).$$

Similarly, if we add and subtract $\mathbb{E}_{\mathcal{D}_{\text{new}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{new}}\right)\right]$ instead of $\mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{new}}\right)\right]$, we have

$$\mathbb{E}_{\mathcal{D}_{\text{new}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{new}}\right)\right] \leqslant \mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{past}}\right)\right] + \mathbb{E}_{\mathcal{D}_{\text{new}}}\left[\ell\left(\tau_{\text{past}}, \tau_{\text{new}}\right)\right] + C\delta(\mathcal{D}_{\text{past}}, \mathcal{D}_{\text{new}}).$$

Combining these results together proves the theorem.

C.3 Randomization and Interference Checks

C.3.1 Randomization Check

We perform covariate balance checks across 274 experiments by calculating the standardized mean difference (SMD), which is the absolute difference in mean covariate values between the treatment and control groups, normalized by the pooled within-group standard deviation. The distribution of SMDs across the experiments is detailed in Table C.1. Notably, out of 10,960 (experiment, covariate) pairs, only 0.3% exhibit SMDs greater than 0.2, and 1% show SMDs exceeding 0.1. This indicates that the randomization process carried out by the focal company was properly implemented.

C.3.2 Testing the No Interference Assumption

This appendix presents empirical evidence for minimal interference caused by multiple offers. Since we observe a correlation between the number of offers a customer receives and their purchasing behavior (likely due to the eligibility criteria specified by the company), we apply the double machine learning framework [40] to examine if the number of additional offers a customer receives causally affects treatment effects.

Specifically, for every offer, we first calculate the doubly robust score $\widetilde{\tau}_{o,i}$ in the same way as in the IRL model. Following this, we apply the debiased machine learning approach to estimate the following partially linear model:

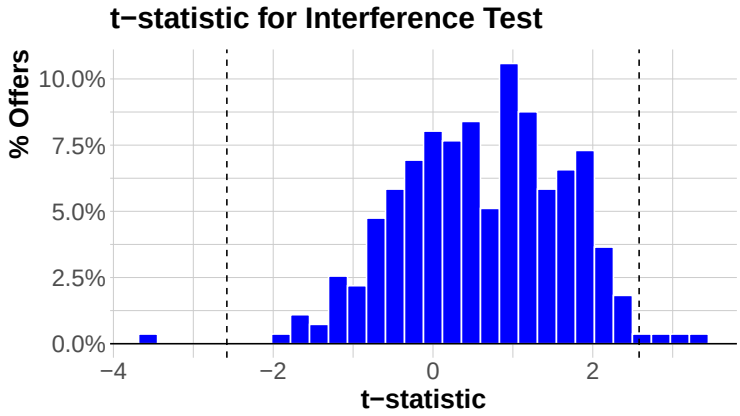
$$\widetilde{\tau}_{o,i} = \beta_o N_{o,i} + m(\mathbf{X}_{o,i}) + \varepsilon_{o,i},$$
$$N_{o,i} = g(\mathbf{X}_{o,i}) + \eta_{o,i},$$

where $N_{o,i}$ denotes the number of effective offers available to customer i when offer o was effective. Here, we use XGBoost with default settings as implemented in the `DoubleML` package [18] with five-fold cross-fitting to estimate the control functions m and g. If interference across offers is absent, we would anticipate that $\beta_o = 0$ for most offers.

Table C.1: *Standarized Mean Differences Across Customer Covariates*

Variable	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
receipt_count	0.001	0.013	-0.051	-0.005	0.007	0.062
aov	0.002	0.012	-0.047	-0.005	0.009	0.046
item_per_order	0.003	0.012	-0.041	-0.003	0.010	0.041
has_past_7	0.001	0.006	-0.016	-0.001	0.004	0.023
has_past_30	0.001	0.004	-0.012	-0.001	0.003	0.014
has_past_60	0.001	0.003	-0.011	-0.001	0.002	0.013
offer_receipt_count	0.034	0.060	-0.066	0.000	0.058	0.340
offer_aov	0.040	0.067	-0.059	0.000	0.067	0.320
offer_item_per_order	0.044	0.075	-0.063	0.000	0.074	0.400
offer_has_past_7	0.008	0.015	-0.022	0.000	0.011	0.069
offer_has_past_30	0.014	0.026	-0.016	0.000	0.021	0.130
offer_has_past_60	0.018	0.034	-0.019	0.000	0.027	0.170
7.ELEVEN	0.000	0.013	-0.051	-0.006	0.007	0.053
AMAZON	0.002	0.015	-0.063	-0.006	0.009	0.058
COSTCO	-0.001	0.012	-0.032	-0.008	0.004	0.034
CVS	0.001	0.015	-0.056	-0.005	0.007	0.056
KROGER	0.001	0.014	-0.038	-0.007	0.009	0.072
MCDONALD'S	-0.001	0.013	-0.056	-0.007	0.005	0.074
TARGET	0.001	0.013	-0.045	-0.007	0.008	0.051
THE HOME DEPOT	-0.001	0.015	-0.063	-0.008	0.007	0.050
WALGREENS	0.002	0.012	-0.039	-0.005	0.009	0.042
WALMART	0.002	0.013	-0.043	-0.005	0.009	0.047
Appetizers & Sides	0.001	0.014	-0.074	-0.006	0.008	0.067
Bakery & Bread	-0.001	0.012	-0.049	-0.007	0.006	0.042
Bath & Body	0.002	0.014	-0.039	-0.005	0.010	0.049
Beer, Wine & Spirits	0.002	0.013	-0.042	-0.005	0.009	0.048
Beverages	0.002	0.013	-0.046	-0.006	0.009	0.055
Candy	0.001	0.013	-0.046	-0.006	0.009	0.048
Cat	0.003	0.014	-0.056	-0.003	0.010	0.051
Dairy	0.000	0.013	-0.057	-0.006	0.006	0.039
Dog	0.001	0.014	-0.039	-0.005	0.008	0.058
Hair Care	0.001	0.014	-0.052	-0.006	0.009	0.046
Household, Paper & Plastic	0.001	0.012	-0.039	-0.006	0.008	0.042
Produce	0.000	0.012	-0.037	-0.007	0.007	0.040
is_referral	0.001	0.007	-0.021	-0.003	0.004	0.037
age	0.000	0.014	-0.046	-0.009	0.009	0.049
tenure	0.002	0.016	-0.052	-0.007	0.012	0.061
is_male?	0.000	0.005	-0.015	-0.003	0.003	0.022
is_femare?	0.000	0.006	-0.023	-0.003	0.004	0.020
is_non_binary?	0.000	0.002	-0.008	-0.001	0.001	0.011

Figure C.1 presents the distribution of t-statistics for the estimated β_o across 274 offers. It is noteworthy that only 1.4% of offers have a significantly non-zero at a 1% significance level, suggesting that potential interference is minimal in our empirical context.



Note. The dashed line indicates the threshold for rejecting the null hypothesis that $\beta_o = 0$ at a 1% significance level.

Figure C.1: *t-statistics for Interference Test across 274 Offers*

C.4 Model Specification for Main Analysis

C.4.1 Doubly Robust Score Estimation

To construct the doubly robust score for each offer, we estimate the nuisance models $\widehat{\mu}_{o,1}^{[-i]}$ and $\widehat{\mu}_{o,0}^{[-i]}$ using XGBoost with ten-fold cross-fitting. For tuning parameters, we utilize the cross-validation function in the R XGBoost package [37] to select the depth of a single tree from $\{3, 4, 5\}$, the learning rate $\eta \in \{0.5, 1, 1.5\}$, and the number of boosting rounds from $\{20, 25, 30\}$. The results suggest that a depth of 3, $\eta = 1$, and 30 rounds yield the best predictive accuracy. However, the results are relatively insensitive to these tuning parameters. For propensity scores, since the treatment assignment is random, we simply use the percentage of customers receiving the offer as the propensity score for each offer.

C.4.2 Deep Learning Models

As described in Section 3.5.1, we implement two five-layer fully-connected neural networks for representation learning and one five-layer fully-connected neural network for prediction. Each

hidden layer contains ten nodes. With the representation $K_1 = K_2 = 10$ the IRL model comprises a total of 1,981 parameters. When $K_1 = K_2 = 50$, the model has 2,401 parameters in total.

For joint models in Section 3.5.2 (i.e., the Deep DR-learners with and without design features), we employ an eight-layer fully-connected neural network. The first hidden layer consists of 20 nodes, while the subsequent layers each have 10 nodes. This configuration results in a total of 2,291 parameters for the Deep DR-learner with design features and 1,491 parameters for the Deep DR-learner without design features. For individual modeling, we employ the same neural network architecture as the deep DR-learner without design features, for each offer.

To calibrate deep neural networks, we use mini-batch stochastic gradient descent with a batch size of 100 for joint modeling and 30 for individual modeling. We adaptively determine the learning rate using the adadelta algorithm [220] implemented in Keras [46]. We train the neural networks for 10 epochs, as the validation performance stabilizes after this training period.

C.5 Targeting Performance with Additional Models

In this appendix, we expand upon the results discussed in Section 3.6.1 by including other benchmark models. We begin by introducing these additional benchmark models and then present key findings for the three targeting scenarios.

C.5.1 Additional Benchmark Models

First, we construct IRL models with varying dimensions of representations, ranging from $K_1 = K_2 = 10$ to 50. Second, for both individual and joint CATE models, we consider the following benchmark models in addition to utilizing deep learning models for predicting the doubly robust scores.

1. **DR-learner with XGBoost**: Similar to the Deep DR-learner, we first calculate the doubly robust score. However, instead of employing a deep learning model, we opt for XGBoost to predict the doubly robust score. We conduct five-fold cross-validation and select a tree depth of 5, a learning rate of 1, and 30 rounds, as these parameters achieve the lowest approximation error for the proxy.

2. **T-learner with Deep Neural Network**: We construct two deep learning-based outcome models: one for treated customers and another for control customers. Subsequently, we determine the CATE predictions by calculating the difference between the predicted outcomes from the treatment model and the control model. We use the same neural network architecture as the individual deep DR-learner and hyper-parameters for model calibration.

Finally, we explore targeting strategies based on predicted purchase amounts, focusing on baseline predictions rather than incremental purchases. To achieve this, we develop a deep neural network (with the same architecture as the joint deep DR-learner) that utilizes offer design features and customer covariates to predict future purchases across training data in each scenario.

C.5.2 Targeting for Existing Offers

Table C.2 presents the average RMSE and AUTOC across 274 offers, replicating the results from Table 3.2. Firstly, the IRL model consistently outperforms other methods, demonstrating superior targeting effectiveness. The performance of the IRL models remains relatively consistent across varying dimensions (rows 1 to 5). Secondly, although the joint DR-learner using XGBoost (shown in row 7) displays relatively strong targeting performance (slightly below IRL, but not significantly), it also has a higher prediction error compared to all the joint deep DR-learners. This result suggests that although the joint DR-learner using XGBoost may be appropriate for the focal firm to use for targeting, it may not be the best tool for evaluating the profitability of a new campaign due to its significantly larger predictive error.

Thirdly, the deep T-learner (listed in row 10) is identified as the least effective, potentially harmful due to its negative average AUTOC. This is likely a result of the high imbalance in the experiments, with only 10% of customers in the control group, leading to a less accurate outcome model for control group customers.[2]

Fourthly, we observe that targeting customers with high predicted purchases is ineffective as it results in negative AUTOCs, indicating that a higher purchase potential does not always correlate with a higher treatment effect. Finally, the individual DR-learner using XGBoost shows better

[2]Such imbalance is common in practice, as firms aim to minimize untreated individuals to conserve resources while maintaining a small control group for estimation purposes.

targeting performance but significantly higher RMSE.

Table C.2: *Predictive Validity for 274 Existing Offers*

Model	RMSE	AUTOC
IRL (10-dim)	7.78 (0.82)	0.77 (0.23)
IRL (20-dim)	8.07 (0.85)	0.77 (0.12)
IRL (30-dim)	7.74 (0.57)	0.78 (0.11)
IRL (40-dim)	7.68 (0.76)	0.82 (0.11)
IRL (50-dim)	7.48 (0.69)	0.74 (0.14)
Joint-DR-DNN	8.29 (1.52)	0.45 (0.20)
Joint-DR-XGBoost	10.9 (1.26)	0.65 (0.10)
Joint-DR-DNN (No Design Features)	8.62 (1.11)	0.48 (0.13)
Joint-t-DNN	15.91 (1.31)	−0.21 (0.11)
High Predicted Purchase	—	−0.10 (0.04)
Individual-DR-DNN	8.57 (1.41)	0.05 (0.20)
Individual-DR-XGBoost	164.42 (32.21)	0.54 (0.11)
Individual-t-DNN	8.27 (0.49)	0.06 (0.21)

Note: We calculated the mean average RMSE and AUTOC values across 274 offers over 20 splits, with standard deviations shown in parentheses.

C.5.3 Targeting for Untested Offers

Table C.3 presents the average RMSE and AUTOC across 55 offers related to personal care products, replicating the results from Table 3.3. Firstly, the IRL models with dimensions $K_1 = K_2 = 10$ and 20 demonstrate better targeting performance compared to models with higher dimensions, where AUTOC tends to decrease as the dimensions of the learned representations increase, underscoring the benefits of dimension reduction for generalization. Secondly, while the Joint-DR-XGBoost (row 7) exhibits the highest RMSE of all models at 10.27, suggesting potential overfitting, it still maintains a moderately good AUTOC. Thirdly, targeting customers with high predicted purchases yields positive AUTOCs, indicating that higher purchase potential correlates with a higher treatment effect for personal care products. Finally, all the individual DR-models underperform compared to the joint models, highlighting the advantages of synergizing information from past experiments.

Table C.3: *Targeting Performance for Offers Related to Personal Care Products*

Model	RMSE	AUTOC
IRL (10-dim)	3.28 (0.56)	0.67 (0.15)
IRL (20-dim)	3.43 (0.41)	0.64 (0.13)
IRL (30-dim)	3.88 (0.43)	0.59 (0.11)
IRL (40-dim)	3.26 (0.36)	0.48 (0.15)
IRL (50-dim)	3.73 (0.39)	0.50 (0.14)
Joint-DR-DNN	3.32 (0.56)	0.31 (0.10)
Joint-DR-XGBoost	10.27 (0.36)	0.52 (0.12)
Joint-DR-DNN (No Design Features)	3.62 (0.39)	0.54 (0.11)
Joint-t-DNN	10.07 (0.35)	0.37 (0.11)
High Predicted Purchase	—	0.31 (0.15)
Actual (Individual-DR-DNN)	5.44 (0.6)	0.02 (0.14)
Actual (Individual-DR-XGBoost)	9.72 (0.59)	0.29 (0.14)
Actual (Individual-t-DNN)	5.95 (0.69)	0.18 (0.16)

Note: We calculated the mean average RMSE and AUTOC values across 55 holdout offers, with standard deviations shown in parentheses.

C.5.4 Targeting for Untested Customers

Table C.4 presents the average RMSE and AUTOC across 31 wine offers targeting customers over 40 years old, replicating the results from Table 3.4. We observe that the IRL models with lower dimensions generally outperform the other models. Additionally, both the joint deep DR-learner and strategies targeting high predicted purchases also show reasonably good targeting performance. However, while the joint model using XGBoost achieved the lowest RMSE, its AUTOC is significantly lower than that of other methods. This suggests that the IRL model, particularly at lower dimensions, provides the most effective robust results and targeting effectiveness.

Table C.4: *Targeting Performance for Wine Offers on Customers with Age over 40*

Model	RMSE	AUTOC
IRL (10-dim)	5.58 (0.77)	1.27 (0.36)
IRL (20-dim)	5.78 (0.61)	1.20 (0.38)
IRL (30-dim)	6.25 (0.78)	1.14 (0.34)
IRL (40-dim)	5.73 (0.79)	1.17 (0.35)
IRL (50-dim)	5.93 (0.72)	1.16 (0.35)
Joint-DR-DNN	5.75 (0.67)	1.16 (0.37)
Joint-DR-XGBoost	4.53 (0.60)	0.70 (0.34)
Joint-DR-DNN (No Design Features)	6.27 (0.83)	0.98 (0.39)
Joint-t-DNN	9.71 (0.66)	0.44 (0.36)
High Predicted Purchase	—	1.11 (0.31)
Actual (Individual-DR-DNN)	6.65 (0.98)	0.15 (0.45)
Actual (Individual-DR-XGBoost)	19.01 (1.42)	0.57 (0.40)
Actual (Individual-t-DNN)	6.97 (1.03)	0.05 (0.45)

Note: We calculated the mean average RMSE and AUTOC values using holdout data, which includes 783,671 observations from 31 distinct offers, with standard deviations shown in parentheses.

C.6 Robustness Check

C.6.1 Depth of Neural Networks

In this appendix, we demonstrate that the predictive validity of the IRL model is robust across various depths of neural network architectures. Specifically, we construct DNN-based models using shallow architectures by removing two hidden layers from each of the neural networks described in C.4. Additionally, we construct models with deep architectures by adding two additional hidden layers (each with ten nodes) to each of the neural networks outlined in C.4.

Table C.5: *Predictive Validity for 274 Existing Offers*

Model	RMSE	AUTOC
Shallow Neural Network Architectures		
IRL (10-dim)	6.90 (1.73)	0.58 (0.21)
IRL (20-dim)	7.51 (0.71)	0.56 (0.19)
IRL (30-dim)	7.69 (0.57)	0.62 (0.13)
IRL (40-dim)	7.62 (0.59)	0.68 (0.14)
IRL (50-dim)	7.33 (0.51)	0.54 (0.16)
Joint-DR-DNN	7.99 (0.80)	0.25 (0.22)
Joint-DR-DNN (No Design Feature)	7.98 (0.45)	0.32 (0.23)
Joint-t-DNN	13.05 (2.50)	−0.24 (0.12)
High Predicted Purchase	—	−0.04 (0.10)
Individual-DR-DNN	9.29 (0.58)	0.12 (0.16)
Individual-t-DNN	13.18 (0.66)	0.15 (0.22)
Medium Neural Network Architectures (Main Analysis)		
IRL (10-dim)	7.78 (0.82)	0.77 (0.23)
IRL (20-dim)	8.07 (0.85)	0.77 (0.12)
IRL (30-dim)	7.74 (0.57)	0.78 (0.11)
IRL (40-dim)	7.68 (0.76)	0.82 (0.11)
IRL (50-dim)	7.48 (0.69)	0.74 (0.14)
Joint-DR-DNN	8.29 (1.52)	0.45 (0.20)
Joint-DR-DNN (No Design Features)	8.62 (1.11)	0.48 (0.13)
Joint-t-DNN	15.91 (1.31)	−0.21 (0.11)
High Predicted Purchase	—	−0.10 (0.04)
Individual-DR-DNN	8.57 (1.41)	0.05 (0.20)
Individual-t-DNN	8.27 (0.49)	0.06 (0.21)
Deep Neural Network Architectures		
IRL (10-dim)	7.22 (0.16)	0.75 (0.16)
IRL (20-dim)	7.17 (0.56)	0.72 (0.17)
IRL (30-dim)	7.10 (0.36)	0.76 (0.14)
IRL (40-dim)	7.57 (2.39)	0.49 (0.32)
IRL (50-dim)	7.04 (0.31)	0.55 (0.11)
Joint-DR-DNN	7.59 (0.45)	0.38 (0.18)
Joint-DR-DNN (No Design Feature)	7.95 (0.54)	0.37 (0.10)
Joint-t-DNN	18.19 (5.73)	−0.11 (0.17)
High Predicted Purchase	25.88 (1.72)	−0.07 (0.05)
Individual-DR-DNN	5.69 (0.26)	0.2 (0.29)
Individual-t-DNN	6.52 (0.38)	0.05 (0.17)
Non-DNN Model		
Joint-DR-XGBoost	10.90 (1.26)	0.65 (0.10)
Individual-DR-XGBoost	164.42 (32.21)	0.54 (0.11)

Note: We calculated the mean average RMSE and AUTOC values across 274 offers over 20 splits, with standard deviations shown in parentheses.

C.6.2 Including Offers with Negative Treatment Effects

In this Appendix, we present the same analysis as in Section 3.6, but this time including offers with negative treatment effects. Table C.6 compares the predictive accuracy and targeting effectiveness across different methods. The key findings are consistent with those in Table C.2, where the IRL model outperforms other methods, although the AUTOC value is smaller and the standard deviation is higher. This is likely because it may not be possible to learn any useful targeting rules for offers with negative treatment effects.

Table C.6: *Predictive Validity for 362 Existing Offers*

Model	RMSE	AUTOC
IRL (10-dim)	7.52 (0.55)	0.38 (0.20)
IRL (20-dim)	7.54 (0.49)	0.38 (0.22)
IRL (30-dim)	7.86 (0.43)	0.40 (0.17)
IRL (40-dim)	7.74 (0.52)	0.37 (0.14)
IRL (50-dim)	7.42 (1.00)	0.34 (0.16)
Joint-DR-DNN	7.76 (0.70)	0.30 (0.18)
Joint-DR-XGBoost	9.16 (0.03)	0.25 (0.14)
Joint-DR-DNN (No Design Feature)	8.79 (0.64)	0.30 (0.26)
Joint-t-DNN	16.00 (0.98)	0.05 (0.22)
High Predicted Purchase	—	0.13 (0.23)
Actual (Individual-DR-DNN)	6.29 (0.22)	0.01 (0.15)
Actual (Individual-DR-XGBoost)	122.37 (1.21)	0.27 (0.10)
Actual (Individual-t-DNN)	8.24 (0.25)	−0.08 (0.20)

Note: We calculated the mean average RMSE and AUTOC values across 362 offers over 20 splits, with standard deviations shown in parentheses.

C.6.3 Additional Example of Targeting for Untested Offers

In this Appendix, we provide an additional example to demonstrate the ability of the proposed IRL method in generalizing to untested offers. Here, we replicate the analysis in Section 3.6.1, but replace the personal care product category with beer and seltzer products as the holdout offers (the largest product category for offers). Table C.7 presents the average RMSE and AUTOC across 57 offers related to beer and seltzer products, replicating the results from Table C.3. The findings are consistent with those for personal care products: the IRL method outperforms all

other methods, and IRL models with lower representation dimensions perform slightly better than those with higher dimensions. The only difference here is that the joint DNN model also performs well in targeting in this situation. Nonetheless, this example highlights the robust performance of the IRL model.

Table C.7: *Targeting Performance for Offers Related to Beer and Seltzer Products*

Model	RMSE	AUTOC
IRL (10-dim)	4.51 (0.3)	0.77 (0.20)
IRL (20-dim)	5.28 (0.72)	0.70 (0.21)
IRL (30-dim)	4.53 (0.31)	0.66 (0.18)
IRL (40-dim)	4.88 (0.34)	0.70 (0.18)
IRL (50-dim)	4.75 (0.35)	0.64 (0.20)
Joint-DR-DNN	5.67 (0.85)	0.65 (0.20)
Joint-DR-XGBoost	20.47 (0.72)	0.44 (0.17)
Joint-DR-DNN (No Design Feature)	5.9 (1.59)	0.57 (0.19)
High Predicted Purchase	—	0.26 (0.21)
Joint-t-DNN	18.32 (8.23)	0.12 (0.22)
Actual (Separate-DR-DNN)	9.87 (5.85)	-0.04 (0.25)
Actual (Separate-DR-XGBoost)	13.4 (0.59)	0.39 (0.17)
Actual (Separate-t-DNN)	9.55 (6.71)	0.08 (0.23)

Note: We calculated the mean average RMSE and AUTOC values across 57 offers over 20 splits, with standard deviations shown in parentheses.

C.6.4 Additional Example of Targeting for Untested Segments

In this Appendix, we provide an additional example to demonstrate the ability of the proposed IRL method to generalize to untested customer segments for certain offers. Here, we replicate the analysis in Section 3.6.1, but now select female customers and offers related to baby products as the holdout set. Table C.8 presents the average RMSE and AUTOC across 52 offers related to baby products, replicating the results from Table Table C.4. The findings are consistent with the main analysis: the IRL method outperforms all other methods, and IRL models with lower representation dimensions perform slightly better than those with higher dimensions.

There are two main differences compared to the results in Table C.4. First, the joint model with traditional deep learning architectures (Rows 6 and 7) performed significantly worse in targeting

performance than the other methods. Second, building individual models using XGBoost on the actual experiment data achieves the best targeting performance (although insignificantly different from IRL models with less than 30 dimensions), while the predictive error of the treatment effects (RMSE) is still significantly higher than IRL. In summary, the IRL model remains the best for decision-making as it achieves the lowest RMSE and high AUTOC.

Table C.8: *Targeting Performance for Baby Offers on Female Customers*

Model	RMSE	AUTOC
IRL (10-dim)	5.83 (0.66)	0.42 (0.23)
IRL (20-dim)	6.00 (0.73)	0.45 (0.28)
IRL (30-dim)	5.90 (0.68)	0.42 (0.21)
IRL (40-dim)	6.08 (0.63)	0.20 (0.28)
IRL (50-dim)	6.40 (0.70)	0.34 (0.28)
Joint-DR-DNN	6.03 (0.74)	−0.15 (0.30)
Joint-DR-XGBoost	6.85 (0.76)	−0.13 (0.25)
Joint-DR-DNN (No Design Feature)	6.61 (0.90)	−0.42 (0.24)
High Predicted Purchase	22.04 (0.91)	−0.22 (0.27)
Joint-t-DNN	—	−0.29 (0.3)
Actual (Separate-DR-DNN)	7.48 (0.76)	0.05 (0.23)
Actual (Separate-DR-XGBoost)	16.82 (1.08)	0.47 (0.25)
Actual (Separate-t-DNN)	7.62 (1.03)	0.10 (0.18)

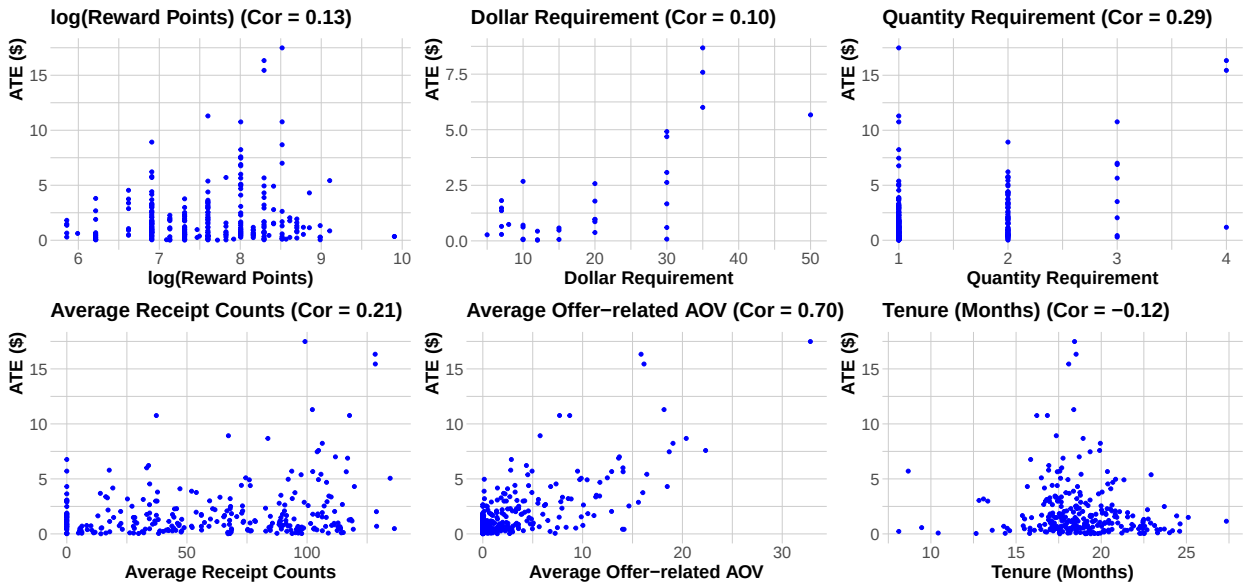
Note: We calculated the mean average RMSE and AUTOC values across 52 offers over 20 splits, with standard deviations shown in parentheses.

C.7 Additional Details for Empirical Results

C.7.1 Correlations between Average Treatment Effects and Observed Vairables

We explore how variations in ATEs correlate with observable characteristics. Figure C.2 depicts the relationship between various design features, the average characteristics of eligible customers for each offer, and their corresponding ATE. In these subplots, the x-axis represents offer attributes, including design features like reward points and customer covariates such as average age. The y-axis measures the ATE for each offer. The patterns observed indicate that (i) both design features and customer covariates provide some predictive value for an offer's ATE,

yet (ii) there are significant variations in ATE, even when specific design features or customer profile variables are accounted for.



Note. Each point reports the average treatment effect of a promotional offer and its corresponding design feature or covariate average.

Figure C.2: *Overview of ATE, Selected Design Features, and Selected Covariate Averages*

C.7.2 Comparison of Personal Care Product Offers with Other Offers

In Section 3.6.1, we assess the targeting performance of personal care offers when the IRL model is trained using offers unrelated to personal care products. This evaluation not only tests the concept shift problem also examines attribute shift. Figure C.3 illustrates the distribution differences for a key design feature (reward points) and a key customer covariate (average receipt counts). The results indicate that personal care offers tend to offer fewer reward points and target customers with higher receipt counts, demonstrating the presence of attribute shifts in this case study.

C.7.3 Offer Clusters Using Raw Design Features

In this appendix, we demonstrate that offer clusters derived from incrementality representations provide a more nuanced understanding than those derived from raw design features. For comparison, we also conduct a similar cluster analysis in Section 3.6.2 using the original

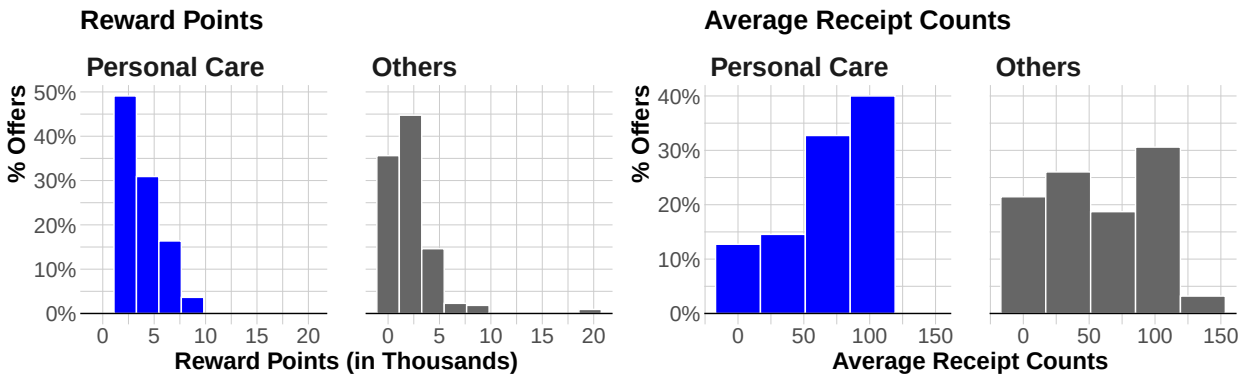


Figure C.3: *Comparison of Personal Care Product Offers with Other Offers*

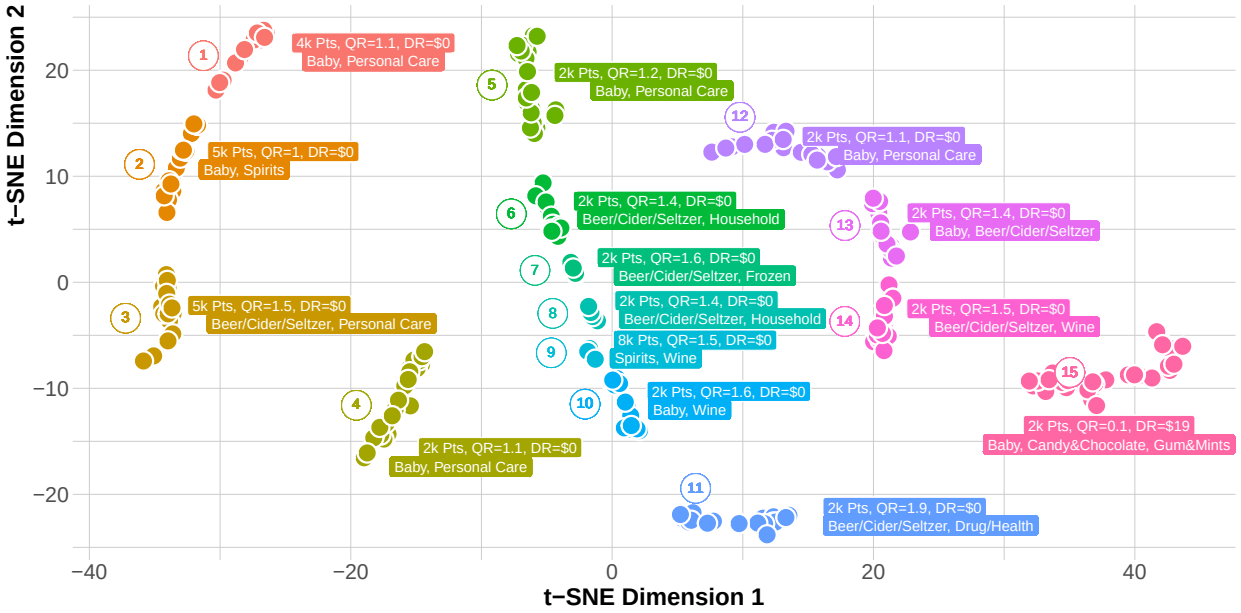
design features ($\mathbf{Z}_o$). We adjust the DBSCAN parameters to obtain a comparable number of clusters, resulting in 13 clusters for the incrementality representations and 15 clusters for the design features. Figure C.4 presents the t-SNE map and cluster results. Unlike the clusters formed using raw features, which primarily differentiate baby and personal care products (clusters 1, 5, and 12 in Figure C.4) based mainly on reward points, the incrementality representation clusters identified in Figure 3.4 provide a more detailed consideration of reward claim requirements.

To further validate that the incremental representation captures more nuanced information, we conduct a statistical analysis to identify the features determining cluster membership. The variable importance plot is shown in Figure C.5. In contrast to Figure 3.5, we find that clusters based on raw design features are overwhelmingly influenced by just two features: reward points, which constitute 80% of the classification outcomes, and dollar requirements, which account for 17% of the outcomes.

C.7.4 Treatment Effect Heterogeneity Using Original Design Features and Customer Covariates

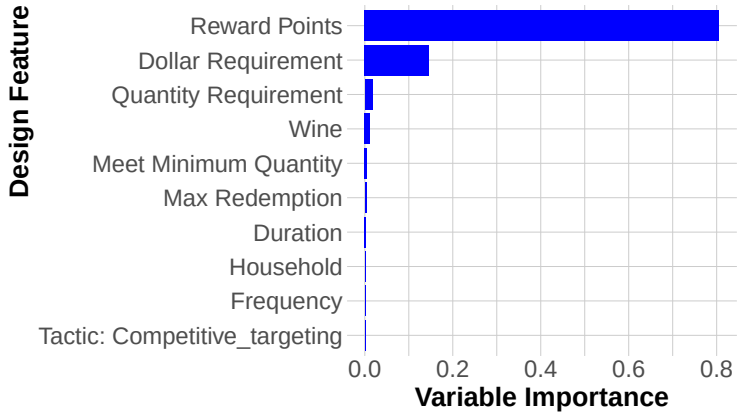
Section 3.6.2 shows that offer and customer clusters based on incrementality representations elucidate clear treatment effect heterogeneity. In this Appendix, we show that clusters based on the original design features and customer covariates cannot explain treatment effect heterogeneity as effectively. Specifically, we replicate Figure 3.6 with offer and customer clusters derived using the original design features and customer covariates.

Figure C.6 shows the GATEs for three clusters associated with 'baby and personal care'



Note. Each point on the map corresponds to a promotional offer, with colors indicating the clusters determined by the DBSCAN algorithm. Labels with colored backgrounds show the average reward points (Pts, set as zero for offers requiring minimum dollar spending), required quantities (QR), required dollar amounts (DR, set as zero for offers requiring certain product quantities), and the top two promoted categories for each cluster.

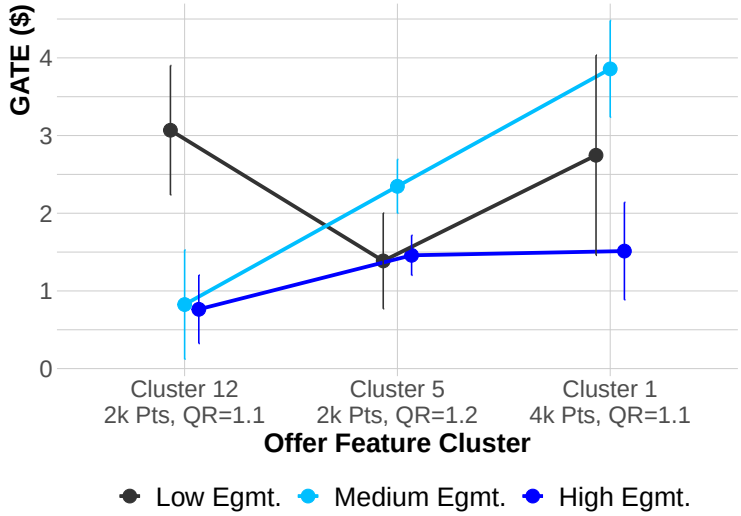
Figure C.4: *t-SNE Map for Offer Design Features ($\mathbf{Z}_o$)*



Note. Variable importance is measured by the average gain in accuracy across all splits where the variable is utilized.

Figure C.5: *Variable Importance Plot of Top 10 Design Features for Cluster Classification Models*

products derived using raw design features (clusters 12, 5, and 1 in Figure C.4)) across three engagement segments identified using (scaled) customer covariates (Table3.5b). The results show minimal differences across defined clusters and segments. In contrast, as depicted in Figure 3.6 in the paper, incrementality representations more effectively elucidate treatment effect heterogeneity. This significant difference highlights the value of key design features and customer segments identified using incrementality representations for tailoring decision making.



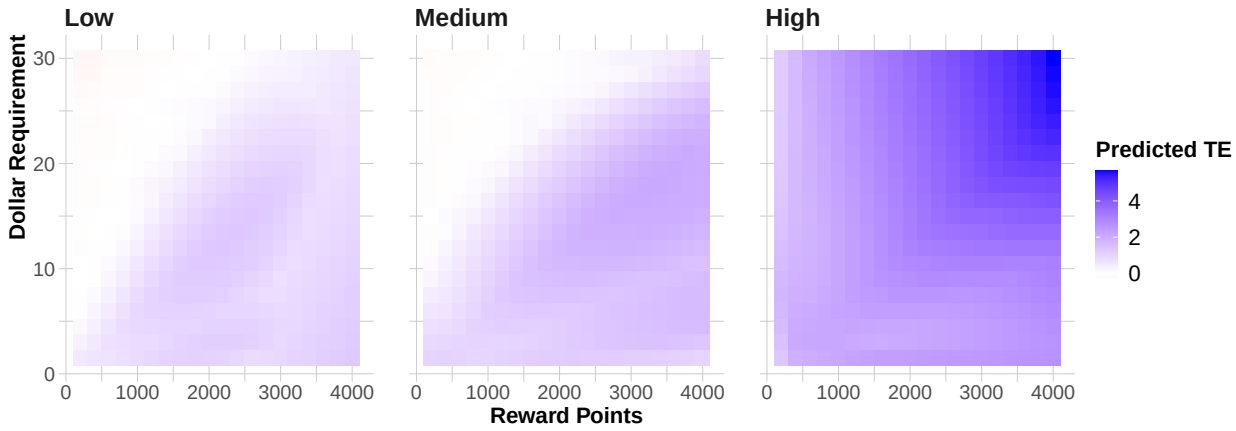
Note. Each point reports the group average treatment effect together with two-standard-error interval. The offer representation clusters align with those outlined in Figure C.4, and the engagement segments correspond to those in Table3.5b.

Figure C.6: *Treatment Effects by Offer Clusters and Customer Segments Derived from* $(\mathbf{Z}_o, \mathbf{X}'_{o,i})$

C.7.5 SIP for Predicted Treatment Effects

In this appendix, we create the SIP of predicted treatment effects on offer-related spending across customer segments identified using incrementality representations. Figure C.7 shows the SIP for the previously described baby product offer, generated using the IRL model. Here, reward points vary within $\{200, \cdots, 4000\}$ and the dollar requirement ranges from $\{\$1.5, \cdots, \$30\}$. The results suggest that promotional designs should be tailored to different engagement segments. For the low engagement segment, offering 2,400 reward points with a minimum \$14 purchase leads to the highest incremental purchase (\$1.35) per customer. For the medium engagement segment, 3,200 reward points with a minimum \$18 purchase maximize incremental purchases (\$4.28) per customer. Additionally, for high engagement customers, 4,000 reward points with a minimum \$30

purchase achieve the greatest incremental purchase ($5.65) per customer.



Note. We set the reward points within the range of $\{200, \cdots, 4000\}$ and the dollar requirement within $\{\$1.5, \cdots, \$30\}$.

Figure C.7: *Segment Incrementality Plot for the Focal Offer Design*