The Roads Not Taken: Model Multiplicity in Machine Learning

HARVARD UNIVERSITY

Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Harvard John A. Paulson School of Engineering and Applied Sciences
have examined a dissertation entitled:

"The Roads Not Taken: Model Multiplicity in Machine Learning"

presented by: Jamelle D. Watson-Daniels

Signature _____
Typed name: Professor David C. Parkes

Signature _____
Typed name: Professor Ariel Procaccia

Signature _____
Typed name: Professor Berk Ustun

Signature _____
Typed name: Professor Alexandra Chouldechova

April 19, 2024

The Roads Not Taken: Model Multiplicity in Machine Learning

A DISSERTATION PRESENTED
BY
JAMELLE D. WATSON-DANIELS
TO THE
SCHOOL OF ENGINEERING AND APPLIED SCIENCES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
APPLIED MATHEMATICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
APRIL 2024

Thesis advisors: David C. Parkes and Berk Ustun Jamelle D. Watson-Daniels

The Roads Not Taken: Model Multiplicity in Machine Learning

Abstract

In machine learning, *model multiplicity* is the existence of multiple models that perform equally well for a given prediction task (also known as the *"Rashomon effect"*). The set of near-optimal models is referred to as the *"Rashomon set." Predictive multiplicity* examines how predictions change over this set of near-optimal models. If model outputs vary significantly across similar models, this information can offer insight into predictive arbitrariness. In this thesis, I introduce frameworks for evaluating and leveraging predictive multiplicity in different settings.

First, I present methods to measure predictive multiplicity in probabilistic classification (predicting the probability of a positive outcome) and develop optimization-based methods to compute these measures efficiently and reliably for convex empirical risk minimization problems. Empirical results show that real-world probabilistic classification tasks can in fact admit competing models that assign substantially different risk estimates. Additionally, I provide insight into how predictive multiplicity arises by analyzing dataset characteristics.

Second, I formulate predictive multiplicity analysis in a resource constrained setting recognizing that predictive allocation tasks are governed by a resource budget. I also extend the multiplicity framing, outlining the concept of *multi-target multiplicity* for quantifying the impact of choices made in regard to target specification for a given predictive allocation task. With this framework, I demonstrate how to fit separate models that are useful for predicting the three outcomes of interest independently and arriving at a way of ranking patients that results in a more equitable allocation.

Third, I investigate the connections between predictive multiplicity and *predictive churn* which is the change in predictions pre- and post- model update in response to a change in training data. I present empirical and theoretical results on characterizing churn in terms of the Rashomon set. Results show that churn unstable points overlap by more than 50 percent with ambiguity points. This points to similarities in the two concepts. Theoretical results to characterize predictive churn between two Rashomon sets as well as churn between models within one Rashomon set hinges on the type of Rashomon set.

I focus on predictive multiplicity to advocate for transparency in the prediction model training procedure. These methods to evaluate predictive multiplicity, as well as connections with predictive churn, contribute to a larger effort for machine learning researchers to be accountable to the individuals affected by model predictions. Similar to a person deciding between roads to take while travelling, insight into alternative options (i.e., roads not taken) may provide insight into the significance of the decisions made.

Contents

Dedicated to my mother, Alisa Renee, whose unwavering love and sacrifices are the bedrock of all my achievements. In loving memory of my grandmother, Alisa Renee Sims, whose guidance and love continues to light my path.

Acknowledgments

At the heart of my journey is my partner in life, Joel O. Anifowose, who has always seen the light in me even at times when shadows clouded my own view. He has been my reviewer, my proofreader, my therapist, my coach - pushing me beyond my imagined limits, my challenger - calling me to stand up in the strength that at times seemed lost, my companion - walking beside me, my dance partner and above all, my friend in joy and hardship. For all the roles you play and for the life we share, I pour out a song of thanks.

I owe so much to my family. My mother (Alisa Renee) was my first teacher in life fiercely advocating for her black female "gifted" child navigating predominantly white schools instilling in me the importance of always remaining grounded in where and who I come from. True to that lesson, my siblings have each played a special part in supporting me on this journey: my sister (Kiana R. Watson), my brother (D'Angelo Watson), my twin brother (Jamal Watson-Daniels), my sister-cousin (Precious Dabbs). To all of you and our larger extended family, I am indebted to your love.

To my son, whose birth was the cornerstone of my journey. Thank you for your lack of concern with anything academic, anything prestigious, anything mathematical, anything reasonable, anything already planned, anything feeding the anxiety in my mind. It has been your freedom and commitment to play that has been the medicine I did not know I needed. Your laughter that made me excited to face each new day. And your birth that awoke something inside me that I needed to survive a global pandemic in isolation with a newborn. For joining me on this journey at precisely the right moment to remind me to "enjoy life", Jaiyeola, I thank you.

A special thank you to my friends and colleagues who have been the emotional backbone during this process. Your community and presence have ensured I always had a listening ear, a shoulder to cry on, gravity to ground me, perspective to help me build my confidence, music to dance, television series to dissect, hope to move forward together. In no particular order, I list a few individuals to whom I express my deepest gratitude: Shannon Whittaker, Jordan Deloach, Yinka Bode-George, Lola Bode-George, Alexx Temeña, Jennifer Chien, Daniel Alabi, Abby Plummer, Mara Freilich, LaNell Williams, Lily Xu, Alana Van Dervont, Alexander Tolbert and many more.

My time at Harvard was enriched by my community at Lowell House where I served as a resident tutor to undergraduates. There will always be a place in my heart for the Lowell House community. To mention a few transformative connections I made, I would like to thank the tutor community, Dean Nina Zipser (faculty dean), Prof. David Laibson (faculty dean), Dean Caitlin Casey (former resident dean), Beth Terry (admin), and all the students who made Lowell feel like home for 3 years.

In preparing this document, I am appreciative of the writing support from Suzanne Smith, the SEAS Graduate Writing Instructor. I am also thankful to Prof. Srijan Kumar and my friends at GA Tech who have graciously welcomed me as a visiting researcher.

Thank you to my Harvard community of academic advisors, collaborators, mentors and more. I came to Harvard in the Applied Physics program where I worked with Prof. Marko Loncar who provided an exceptional community of researchers with whom I have built lasting relationships with. With his full support, Marko encouraged me to follow my gut and make the switch into machine learning. I am grateful. As I became acquainted with literature on algorithmic fairness, Prof. Yiling Chen provided guidance and support through my transition from physics. There are many nuggets of advice she offered that I carry with me to this day. Prof. David C. Parkes became my main Harvard advisor and has been the best guide and example I could have asked for. His unmatched work ethic, integrity, dedication to rigor and commitment to students have been invaluable to me. It has been an honor to know him, to be trained by him and to have learned so much from him.

To my co-advisor, Prof. Berk Ustun. Beyond being an advisor, Berk has spent hours that turned into days that turned into weeks helping me learn how to write a full research publication from start to finish. He has pushed me to become a better writer and researcher always challenging me to consider to larger narrative and story within my work. His insight and dedication to my growth have left an immeasurable influence on my foundation and I am beyond thankful to have had him on this journey.

A big thank you to Prof. Alexandra Chouldechova and Prof. Ariel Procaccia who have reviewed these chapters in grave detail on numerous occasions providing feedback and suggestions to guide progress in the completion of this dissertation. I want to underline that Alex Chouldechova has gone above and beyond as a co-author and mentor. Her expertise in mathematics has greatly improved my work. It is often said that we cannot be what we cannot see. In working alongside Alex, I had the privilege of witnessing her balancing motherhood and scientific research (seemingly) flawlessly. Having the opportunity to work closely with and be mentored by her has enabled me to envision converging on a balance for myself.

I come from people who paid a high price for freedom. People who could only dream of a promised land of milk and honey or institutions where wealth and privilege flow freely. I give thanks and honor to my ancestors for their strength and courage in that fight.

Most recently, my grandmother passed away in July 2023; I am admittedly still in the throws of grief as I write this. It is difficult to overstate the central role she has played in my life and on this journey in particular. Growing up, she stepped in as my mother's secondary co-parent to help raise me. Since beginning my PhD in 2018, almost every conversation ended with her reminding me that she cannot wait to travel to Harvard one day in celebration of me finishing. I never imagined that she would not physically be here to see that day. But I know that grief is one of the most common threads that connects us as humans who have known love. And I know she is here with us in spirit.

To my Mau Mau, I write to you from the promised land, from the land of milk and honey, from the ivory tower where eyes looked down on you, from atop the mountain you weren't allowed to climb just to say: I made it. You made it. We made it.

Citation to Previously Published Work

1. A significant portion of the work presented in Chapter 2 is based on the following publication: Watson-Daniels, J., Parkes, D. C., & Ustun, B. (2023b). Predictive multiplicity in probabilistic classification. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23: AAAI Press

2. A significant portion of the work presented in Chapter 3 is based on the following publication: Watson-Daniels, J., Barocas, S., Hofman, J. M., & Chouldechova, A. (2023a). Multi-target multiplicity: Flexibility and fairness in target specification under resource constraints. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23 (pp. 297–311). New York, NY, USA: Association for Computing Machinery

3. A significant portion of the work presented in Chapter 4 is based on the following working paper: Watson-Daniels, J., du Pin Calmon, F., D'Amour, A., Long, C., Parkes, D. C., & Ustun, B. (2024). Predictive churn with the set of good models. arXiv:2402.07745

The function of freedom is to free someone else.

Toni Morrison

1 Introduction

Artificial intelligence and machine learning (ML) touch most aspects of modern life. Specifically, ML prediction problems have become pervasive across many domains of decision-making[89]. In these instances, predictions about future outcomes are used to influence policy decisions, resource allocation or intervention strategies. From education to health, ML predictions can affect the lives of real people[153,17]. In healthcare, models assign predictions that inform treatment decisions[122,160,85]. In consumer finance, lenders use model predictions to underwrite loans[5,9]. In crimi-

nal justice, models assign predictions that guide sentencing and parole decisions[6,102].

With the widespread use of ML, the discipline of algorithmic fairness has been catalyzed to focus on potential biases in predictive models. In earlier work, Dwork et al. [50] introduced a framework for studying fairness in classification from the perspective of individuals. This laid the groundwork for what is known as *individual fairness* where individuals who are similar should receive similar outcomes[50,81,99]. On the other hand, work on *group fairness* is concerned with some notion of statistical parity for members of subgroups stratified by protected attributes (race or gender)[29,34,91,135] The distinction between individual and group fairness resulted in much discussion about the trade-offs and assumed conflict between the two types of fairness[16,116].

In addition to developing methods in fair classification, research on how algorithms affect people's lives is also concerned with accountability[15,178,44] and transparency[141,45,2]. In terms of accountability, one line of research focuses on the *recourse* available to an individual receiving a prediction by asking what actions an individual could take to change their classification outcome[164]. In a sense, ensuring actionable recourse[164] is seen as one way to hold the model accountable to the individuals affected by that model.

Along with providing individuals with guidance as to the kinds of changes they could make to change a classification outcome, researchers can ask whether decisions made by model developers could change individual classification outcomes. In this questioning, ML researchers and engineers can be accountable to the individuals affected by the model by being more transparent about the ML model development process. To advocate for transparency in the prediction model training procedure, an interesting question is: *Are there multiple equally good ML models that would have changed this individual's prediction?* Said another way, *What are the roads one could have taken to arrive at the prediction model now used to inform individual decisions?*

Consider a person deciding between roads to take while travelling. In terms of the fastest route, there might be one road that beats the others. Or there may be two roads with equal travel time,

in which case, other factors need to be taken into consideration like whether there is a gas station on the route or the likelihood of increased traffic or preference for highway versus back roads. The point is that insight into these options is valuable to the traveler while deciding which road to take. In our context, the ML researcher or engineer becomes the traveler. Further, transparency into the decision making process can help an outsider evaluate the significance of the decision, which can be quantified by examining available alternative options.

In terms of decisions made during ML prediction problem formulation [133], the consequences can be characterized by understanding what changes over possible alternative model selection. For instance, multiple models with near-optimal performance for a given prediction task can exist [25]. If model outputs vary significantly between these similar models, then choosing one model over another has relative importance. Further, the model selection decision might come under scrutiny if a similar model with better fairness properties could have been selected [20]. Even long before deployment, when translating high-level goals into tractable predictive tasks, many reasonable target variable options may be worth considering [133]. Once again, target specification becomes especially high stakes when one target leads to more disparate treatment than other options.

In ML, *model multiplicity* is the existence of multiple predictive models that perform equally well for a given prediction task (also known as the *Rashomon* Effect) [25]. The term *Rashomon* comes from a popular Japanese movie called Rashomon from 1950 where characters provide contradictory reports of the same incident [98]. For ML prediction problems, the set of near-optimal models with similar performance is referred to as the *Rashomon* Set. *Predictive multiplicity* examines how predictions change over the *Rashomon* Set [112].

Analyzing predictive multiplicity enhances transparency into the predictive arbitrariness of a given training procedure, which can help calibrate trust in model predictions among stakeholders interacting with the model [82]. For instance, clinical professionals are increasingly incorporating ML into prediction tasks [122,160,85]. If there are individuals or patients whose prediction changes over the

Rashomon Set, then, medical experts can use this information to take a second look at these patients. In consumer finance, lenders use predictive models in support of predicting the likelihood that a borrower will fail to make payments or default on a loan [5,9]. Consider a consumer who is denied a loan based on the ML model output. If this individual's prediction changes over the *Rashomon* Set, then that decision may be considered arbitrary and hard to justify [22]. Ultimately, if end-users are informed about predictive multiplicity, they could abstain from prediction [21,69], defer a decision to a human expert [124,94] or otherwise readjust their reliance on model outputs.

Given this compelling motivation and the timely need for transparency, predictive multiplicity is the main focus of this dissertation. In the next three chapters, I introduce frameworks for evaluating and leveraging predictive multiplicity in different settings. The main contributions in each chapter are described below.

Chapter 2: Predictive Multiplicity in Probabilistic Classification

Probabilistic classification is often incorporated into real-world risk assessment tasks to inform decisions. For instance, probabilistic classifiers that predict consumer default risk are used by lenders to underwrite loans. We have developed a framework for investigating predictive multiplicity in this setting. More precisely, predictive multiplicity is the prevalence of *conflicting* predictions over the Rashomon set of near-optimal models [112]. For predictive multiplicity analysis with respect to a baseline, we begin by training an optimal *baseline* model which is the solution to an empirical risk minimization (ERM) problem. In probabilistic classification, the baseline model assigns a risk estimate to each individual in the sample i.e. training dataset. We say that a risk estimate assigned by a model in the Rashomon set is *conflicting* if it differs from the baseline risk estimate by at least some deviation threshold. Our aim is to extend the predictive multiplicity measures introduced by Marx et al. [112] for binary classification to probabilistic classification.

Marx et al. [112] define *ambiguity* and *discrepancy* as measures of predictive multiplicity. *Ambiguity* is the proportion of individuals assigned conflicting predictions over the Rashomon set of

near-optimal models. *Discrepancy* is the maximum proportion of individuals assigned conflicting predictions by a single model in the Rashomon set. Measuring multiplicity in probabilistic classification is complicated by the need to clarify the meaning of *conflicting*. In effect, what constitutes a conflicting risk prediction can change across applications (e.g., predictions that vary by 5% or 30%). Likewise, what constitutes a near-optimal model can change across applications depending on how the Rashomon set is defined. This chapter addresses both of these problems by introducing methods that allow users to specify near-optimal metric when defining the Rashomon Set and determine what is meant by *conflicting* (deviation threshold).

To this end, we consider loss, accuracy, and calibration error as possible near-optimal metrics and redefine the two measures, ambiguity and discrepancy, in this setting. We also introduce the *viable prediction range*, which captures how individual predictions change over the Rashomon set. The *viable prediction range* is the smallest and largest risk estimate assigned to an example over competing models in the Rashomon set.

Our optimization-based methods compute our measures reliably. To compute ambiguity and viable prediction ranges, we construct a pool of *candidate models* that assign a specific risk estimate to each example. We train each of these models by solving a constrained convex optimization problem. From these models, we select those with performance within ε of the baseline model as the set of competing models. Specifically, for each threshold probability $p \in P$, we train a candidate model such that the probability assigned to the example is constrained to the threshold p. Marx et al. [112] use a similar *candidate model* approach but for 0-1 loss which is not immediately transferable to our setting where we work with logistic regression.

To compute discrepancy, we formulate a mixed-integer non-linear program (MINLP), which involves constructing a linear approximation of the loss using an iterative, outer-approximation method to solve. This method is exact for computing discrepancy in terms of near-optimal loss. For other metrics, we can again treat the intermediate solutions to the outer approximation algorithm

as candidate models and use these candidates to recover a lower bound similar to the method used to compute ambiguity and viable prediction ranges. Marx et al. [112] follow a similar approach, but in their case they compute discrepancy by solving a mixed integer program (MIP) rather than an MINLP, where their MIP minimizes agreement while constraining the output model to be near-optimal with respect to 0-1 loss.

Via systematic experiments on synthetic data, we offer insights into why predictive multiplicity arises. We find that predictive multiplicity is more prevalent for examples that are both outliers and close to the discriminant boundary, for datasets that are less separable, and for minority groups when a dataset has a majority-minority structure.

Lastly, we present an empirical study on seven real-world risk assessment tasks: risk that a mammogram shows breast cancer [52], risk that a customer will default on a loan [182], risk that a person opens a bank account after a marketing call [123], risk that a person earns over $50,000 [93], risk of rearrest for a crime [3], risk a patient will be diagnosed with obstructive sleep apnea [165]. We show that probabilistic classification tasks can in fact admit competing models that assign substantially different risk estimates. For one set of selected near-optimal error tolerance, results show ambiguity values of 35.3% (breastcancer), 95.8% (sleep apnea), and 51.4% (rearrest). For intuition, this means that 35.3% of breast cancer risk estimates vary by at least 20% over near-optimal models. Results also show discrepancy values at 3.6% (breastcancer), 1.2% (sleep apnea), and 5.4% (rearrest).

Our results also demonstrate how multiplicity can disproportionately impact marginalized individuals. For example, when analyzing predictive multiplicity for the task of predicting the risk of rearrest, individuals who are ethnically Hispanic are disproportionately affected by predictive multiplicity: ambiguity is 39% for African Americans and 49% for Caucasians, compared to 98% for Hispanics. Hence, reporting predictive multiplicity at the subgroup level can also reveal valuable insights.

Chapter 3: Multi-Target Multiplicity: Flexibility and Fairness in Target Specification un-

der Resource Constraints

As noted above, prediction problems are ubiquitous across many domains of decision-making, from employment, to education, to health [89]. Yet real-world problems rarely present themselves as fully formed machine learning tasks [139]. Critically, it is often not clear what target should be predicted to help decision makers achieve their goals [70,133]. It is far from obvious, for example, how employers should go about making such choices in their hiring practices: if the goal is to hire the *best* people, what exactly should the model be predicting [8,133,87]? For a sales position, employers might choose to predict annual sales figures. But they could alternatively choose to predict how well the applicant will work with others, whether customers will actually enjoy interacting with the applicant, etc. Even in domains where target choice might seem more obvious, there can still be a good deal of flexibility in this choice.

A recent line of work has explored the implications of this flexibility in target variable choice for algorithmic fairness considerations. In particular, researchers have pointed out that different choices for the prediction target can lead to more or less disparity in selection rates across groups [133,129,77,121,125,111,87,55]. One particularly well-known study by Obermeyer et al. [129] illustrates both the risks and benefits of target choice. The authors examine an algorithm developed by a healthcare system used in determining patient eligibility for a high-risk coordinated care management program. They find that the healthcare system's choice to adopt healthcare costs as the target of prediction led to notable and *avoidable* racial disparities. Because Black patients in the United States generally incur lower health care costs at equal levels of underlying health care needs, predicting costs results in a score that systematically prioritizes healthier White patients over less healthy Black patients. The authors show that a good deal of the racial disparity could have been avoided had the healthcare system instead chosen to predict a more direct measure of health outcomes. The study has been received as a important lesson in the dangers of insufficiently careful target choice. But it also highlights that practitioners can take advantage of the latitude afforded by target choice

to reduce selection rate disparities.

Though existing literature offers several domain specific examples that highlight the potential importance of target variable choice, prior work does not offer a more general mathematical or computational framework for characterizing the extent to which target variable choice affects individuals' outcomes and selection rate disparities across groups. This work aims to fill this gap. Specifically, we draw connections to recent work on predictive multiplicity. By analogy, in the motivating *multi-target* setting—where there are many possible reasonable prediction targets to choose from—we can consider the set of "good models" that arises from models that predict any of the individual targets well, models that predict a combination of targets well, or that combine the predictions of single-target models. We formalize these ideas and provide examples in the technical sections of the chapter.

As the main contribution, we borrow the "multiplicity" framing to outline the concept of "multi-target multiplicity" for quantifying the flexibility in target specification for a given predictive allocation task: tasks where historical data is used to learn a "prioritization" or "risk" score and that score serves as the basis for deciding how to allocate resources. We introduce the concept of multi-target multiplicity alongside a framework for assessing multi-target multiplicity for predictive allocation tasks. We demonstrate how the framework can assess fairness-related measures by presenting a MIP that calculates the minimum and maximum attainable selection rate for a given group. We demonstrate our framework on the healthcare dataset released by Obermeyer et al. [129]. As expected, results replicate the original result where modeling active chronic conditions produces the highest concentration of current illnesses in the high-risk set and there is more than a 10 percent difference in the racial composition of the high-risk set. We show that the optimized composite model does a reasonable job of capturing each of the individual targets that it is comprised of, but also produces a high-risk set with a high concentration of Black patients. In essence, we are able to fit separate models that are useful for predicting the three outcomes that are of interest on their own

while also arriving at a way of ranking patients that results in a more equitable allocation of a scarce resource using the proposed framework.

Then, we use semi-synthetic data to gain a better understanding of the conditions for which we might (or might not) expect to see such gains in other datasets. The demonstration shows that the optimized composite model can learn to average out unhelpful correlation structure between the protected attribute and the target variables.

Our secondary contribution is to extend concepts from Chapter 2 to *predictive allocation tasks*. There is only a finite amount of benefit, burden, or scrutiny that the system is able to allocate. For instance, the algorithm investigated by Obermeyer et al. [129] was developed to help allocate coordinated care management to a certain number of clients. Similarly, employers cannot offer jobs to *everyone* they predict will be a sufficiently good employee, whatever target or set of targets they choose to predict. Given their limited budgets, they are likely only able to offer jobs to a select few applicants. This means that analyzing a set of good models might need to involve considerations for changes in resource allocations in addition to classification outcomes.

Therefore, we formulate predictive multiplicity in the presence of decisions under resource constraints. Recall, prior work to compute ambiguity involves constructing a pool of candidate models that change individual predictions. From that pool of models, those with near-optimal performance are selected to compute ambiguity. These methods are indirect in that the optimization does not directly constrain these candidate models to be within the Rashomon set. The previous models are formulated to minimize loss such that individual predicted probability is constrained to deviate. Under resource constraints, we develop a MIP that **does** include a constraint on the model performance. The constraint to produce a model within the Rashomon set involves theoretically showing how to include a constraint that neatly characterizes the Rashomon set for linear regression models. In this context, we say an individual instance is flippable if the selection decision (whether that individual will be selected to receive the resource) changes between being selected to not being se-

lected or vice versa. Additionally, we show theoretically that (i) one can efficiently determine that many points are provably *not flippable* over the Rashomon set, and (ii) one can identify a subset of *flippable* points by solving a proxy optimization problem with a closed-form solution that produces a model within the Rashomon set that *may* flip some points into the top. This means that, in practice, we only need to solve the computationally expensive MIP for a small subset of points whose flippability remains undetermined. This methodological improvement to methods in Chapter 2 is enabled by the resource constrained setting.

Chapter 4: Predictive Churn with the Set of Good Models

One of the foremost challenges faced in the deployment of machine learning (ML) models used in consumer-facing applications is unexpected changes over periodic updates. Model updates are essential practice for maintaining and improving long-term performance in mass-market applications like recommendation and advertising. In applications like credit scoring and clinical decision support, however, changes in individual predictions may lead to inadvertent effects on customer retention and patient safety. Consider an individual applying for a loan or a patient being considered for a high-risk treatment program. It may be problematic if their approval or selection decision hinges on whether they applied before or after a model update.

Unexpected or unreliable predictions after an ML model update can illicit safety concerns when models influence human decision-making. Here, predictive instability after a model update can lead to, say, loan denials to applicants who previously would have been approved – even if the new model is more accurate on average. Hence, this chapter focuses on bridging together two facets of predictive (in)stability in applied ML: *predictive churn* and predictive multiplicity.

Predictive Churn considers the differences in individual predictions between models pre- and post-update, where the update is triggered by a change in training data. Predictive churn is formulated in terms of two models: a current model, and an updated model resulting from training the current model on additional fresh data [38]. In several applications, a high level of predictive churn is

undesirable.

The goal is to examine the relationship between predictive churn and predictive multiplicity. In this chapter, the Rashomon Set considered is mainly an empirical Rashomon Set resulting from varying random seed initialisation in training a deep neural network (DNN). This conceptualization of the Rashomon Set has been adopted in prior work [105,51]. First, we examine whether individual predictions that are unstable under model perturbations (predictive multiplicity) are also those that are unstable under dataset perturbations (predictive churn). For a fixed test sample, we find that the set of ambiguous examples does often contain most examples that churn over an update. In practice, analyzing predictive multiplicity for a model could help anticipate examples that will churn over future model updates.

Next, we theoretically characterize the expected churn between models within the Rashomon set from different perspectives. First, we derive an upper bound on the churn between an optimal baseline model and a competing model within the Rashomon set. Then, we show that the expected churn difference between two models within an empirical Rashomon set is zero when we operate without an optimal baseline and under a randomized training procedure. This analysis reveals that the potential for reducing churn by substituting the current deployed model with an alternative from the Rashomon set hinges on the training procedure employed to generate said Rashomon set. It also implies that if future updated models are hypothetically confined to be with the Rashomon set (with respect to a baseline), then the expected churn will be bounded. And finally, operating under the premise that we only have access to Model A, we analyze whether one model within the Rashomon set might result in less churn compared to another model within the Rashomon set. To examine the expected churn difference between any two models within the ε-Rashomon set, we derive an upper bound on the expected churn difference between two Rashomon sets with respect to an optimal baseline model. The results also show that when updating from model A to model B, we can produce both Rashomon sets and analytically compute an upper bound on the churn

between them. Again, the feasibility of reducing churn by substituting the current model with an alternative from the Rashomon set depends the type of Rashomon set.

Finally, we present empirical results on two model types: a standard DNN and an uncertainty aware DNN. In particular, we implement a technique that is common in industry settings *Spectral-normalized Neural Gaussian Process* (SNGP)[104]. Multiplicity and churn are defined in the same way for both kinds of models, but we are interested to understand whether models with inherent uncertainty quantification abilities might (i) exhibit less predictive multiplicity and (ii) whether the uncertainty estimates for a prediction can be predictive as to which examples will be ambiguous, or churn over model updates. Our findings show that in fact there can be more predictive multiplicity for an uncertainty aware (UA) model, though the uncertainty estimates do prove helpful in anticipating unstable instances from the perspective of both predictive multiplicity and churn.

We have been raised to fear the yes within ourselves, our deepest cravings. But, once recognized, those which do not enhance our future lose their power and can be altered. The fear of our desires keeps them suspect and indiscriminately powerful, for to suppress any truth is to give it strength beyond endurance. The fear that we cannot grow beyond whatever distortions we may find within ourselves keeps us docile and loyal and obedient, externally defined, and leads us to accept many facets of our oppression as women.

Audre Lorde

2

Predictive Multiplicity in Probabilistic Classification

2.1 Introduction

Probabilistic classification is often incorporated into real-world risk assessment tasks to inform decisions. For instance, probabilistic classifiers that predict consumer default risk are used by lenders

to underwrite loans[9,5]. Similarly in clinical applications, physicians make treatment decisions using models that predict whether a person suffers from a serious illness[160,85,31]. In criminal justice, judges often make parole and sentencing decisions guided by models that predict the probability that a person will fail to appear in court[6,102,35,183].

The standard approach to selecting a probabilistic classifier often involves optimizing a loss function via empirical risk minimization. But for a given prediction task, there may exist multiple models that perform almost equally well, which is referred to in machine learning as model *multiplicity*[25]. These near-optimal, *competing models*, have similar performance but characteristic differences - e.g. their interpretability[147], explainability[56,47], counterfactual invariance[51], or fairness[40,19,1]. These differences can drastically change how we develop, choose, and use models[22].

We investigate how predictions change across competing models by studying *predictive multiplicity*: the prevalence of conflicting predictions over competing models[112]. To understand our motivation, consider the significance of competing models assigning vastly different predictions in practice. In mortality prediction, a conflicting risk prediction would alter treatment decisions and health outcomes[122]. In drug discovery, a conflicting risk prediction could switch the compounds chosen for confirmatory experiments[158]. By measuring and reporting the prevalence of conflicts, we can improve how we choose and use machine learning models. If end-users know that an individual risk estimate conflicts over the set of competing models, they could abstain from prediction[21,69] or defer a decision to a human expert[124,94]. These implications underline the importance of measuring and reporting predictive multiplicity more widely.

Our main contributions are:

1. We introduce measures of predictive multiplicity in our setting. The Viable Prediction Range examines how multiplicity affects predictions. Ambiguity and discrepancy reflect the proportion of individuals assigned conflicting risk estimates by competing models.

2. We develop optimization-based methods to compute our measures for convex empirical risk minimization problems. This includes employing mixed-integer non-linear programming and outer-approximation algorithms. Whereas previous work defines competing models over a single performance metric, our methods enable developers to examine additional near-optimal metrics.

3. We offer insights into why predictive multiplicity arises via systematic experiments on synthetic data. We find that predictive multiplicity is more prevalent for examples that are both outliers and close to the discriminant boundary, for datasets that are less separable, and for minority groups when a dataset has a majority-minority structure.

4. We present an empirical study on seven real-world risk assessment tasks. We show that probabilistic classification tasks can in fact admit competing models that assign substantially different risk estimates. Our results also demonstrate how multiplicity can disproportionately impact marginalized individuals.

2.2 Related Work

Our work is positioned alongside research on *model multiplicity*. This effect has been referenced in the statistics literature. For example, Chatfield [30] calls for performing a sensitivity analysis over competing models, while Breiman [25] cites multiplicity as a reason to avoid explaining a single model to draw conclusions about the broader data-generating process. Recent advances in computation make multiplicity analysis possible, leading to a stream of research on how competing models differ [56,47,147,51,168,134,40,19,1]. And there are growing discussions about the policy implications of the existence of multiple equally good models [22,20].

There is a growing body of research focused on Rashomon sets. Dong & Rudin [47] explore the variability of feature importance across models in the Rashomon set and propose a method to help

understand and visualize feature importance stability. Semenova et al. [148] examine the existence of simpler models within the Rashomon set underlining that Rashomon sets can be leveraged to search for simpler and more interpretable models. Xin et al. [180] focus on exploring the entire Rashomon set of sparse decision trees which contributes understanding of diversity and quality of models within the Rashomon set, specifically in the context of sparse decision trees. Liu et al. [103] present a method for fast and accurate generation of interpretable risk scores from Rashomon sets with an emphasis on interpretability. Wang et al. [173] introduce an interactive tool for exploring sparse decision trees from Rashomon sets which provides a user-friendly interface to aid in model selection. Wang et al. [172] also introduc an interactive tool for generalized additive models within Rashomon sets with the goal of improving model interpretability for this particular type of Rashomon set. Semenova et al. [146] show that introducing controlled noise into data can lead to simpler models within the Rashomon set which further underlines how Rashomon sets can support interpretability. Zhong et al. [185] go deeper into Rashomon sets with sparse Generalized Additive Models presenting tools for finding more interpretable model within this set. Donnelly et al. [48] introduce the Rashomon Importance Distribution for characterizing variable importance across the Rashomon set which provides a more robust understanding of variable importance.

Similarly, research on multiplicity from the perspective of predictive arbitrariness has been on the rise. Creel & Hellman [42] examine algorithmic decision-making systems and argue that deployment often embodies a form of governance termed *algorithmic leviathan* which draws from a metaphor in political philosophy. Long et al. [105] explore the fairness-accuracy tradeoff in the context of predictive multiplicity. Researchers also investigate the broader impact on group fairness from the perspective of predictive arbitrariness [36] and randomness [60]. Also, Kulynych et al. [97] connect predictive multiplicity directly to differential privacy demonstrating predictive multiplicity increases with the level of privacy.

Our work is distinctly focused on how multiplicity affects prediction. Our approach builds on

Marx et al. [112], who study this effect in classification tasks with yes-or-no predictions. As shown in Figure 2.1, their measures and methods do not extend to our setting. We need definitions that consider risk estimate change as opposed to just yes-no prediction change. Measuring multiplicity in probabilistic classification is complicated by the need to clarify the meaning of "conflicting". In effect, what constitutes a conflicting risk prediction can change across applications (e.g., predictions that vary by 5% or 30%). Likewise, what constitutes a "competing" model can change across applications. The present work addresses both of these problems by introducing methods that allow users to specify what is "competing" (near-optimal metric) and what is "conflicting" (deviation threshold). Also, previous work has yet to examine why predictive multiplicity arises, which we contribute to. Note that Hsu & Calmon [76] introduce a metric called *Rashomon Capacity*, which quantifies score variations among models in the Rashomon set in the setting of probabalistic classification. Whereas we focus on extending existing measures to this setting, Hsu & Calmon [76] focuses on adding an additional metric to the literature.

One way we compute predictive multiplicity is by constructing a range of individual risk predictions as a way to quantify pointwise uncertainty resulting from an underspecified empirical risk minimization problem. This relates to methods for evaluating predictive uncertainty such as conformal prediction [150,143] as well as Bayesian approaches see e.g., [49,106]. However, conformal prediction focuses on uncertainty that arises due to non-conformity between historical data and new data, which is orthogonal to our goal. We focus on a non-Bayesian approach, recognizing that non-Bayesian methods are very typical in applied machine learning. Our goals relate also to a line of work that aims to quantify and communicate uncertainty in machine learning [74,83,114,157,94?] and calibrate trust among stakeholders [82]. Other complementary work seeks interventions to resolve multiplicity [1] or ensembling [21].

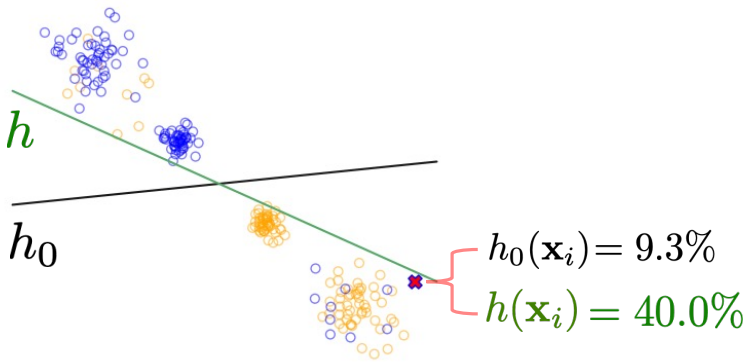


Figure 2.1: Classification models that make the same yes-or-no predictions can still assign conflicting risk predictions. Here, we show a 2D classification task with $n^+ = 200$ positive examples (blue) and $n^- = 200$ negative examples (orange). We plot the decision boundary of a baseline model h_0 (black; log-loss/AUC/calibration = 0.41/0.88/17%) and a competing model that performs almost equally well $h(\boldsymbol{x}_i)$ (green; log-loss/AUC/calibration = 0.42/0.89/16%). As shown, both classifiers make the same yes-or-no predictions, but assign conflicting risk estimates to individual examples e.g., example $\boldsymbol{x}_i$ is assigned a risk estimate of $h_0(\boldsymbol{x}_i) = 9.3\%$ by the baseline model but $h(\boldsymbol{x}_i) = 40.0\%$ by the competing model.

2.3 Framework

We consider a probabilistic classification task with a dataset of n examples $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$. Each example consists of a feature vector $\boldsymbol{x}_i = [1, x_{i1}, \ldots, x_{id}] \in \mathcal{X} \subseteq \mathbb{R}^{d+1}$ and a label $y_i \in \mathcal{Y} = \{-1, +1\}$, where $y_i = +1$ is an event of interest (e.g., default on a loan). With the dataset, we train a probabilistic classifier $h : \mathcal{X} \to [0, 1]$ – i.e., a model that assigns a risk estimate to example $\boldsymbol{x}_i$ as: $h(\boldsymbol{x}_i) := \Pr(y_i = +1|\boldsymbol{x}_i)$. We refer to this model as the *baseline model*, h_0, because it is the optimal solution to an empirical risk minimization (ERM) problem of the form:

$$\min_{h \in \mathcal{H}} L(h; \mathcal{D}), \tag{2.1}$$

where $\mathcal{H}$ is a family of probabilistic classifiers, and $L(\ \cdot\ ; \mathcal{D})$ is a loss function evaluated on the dataset $\mathcal{D}$. In what follows, we write $L(h)$ instead of $L(h; \mathcal{D})$ for conciseness. We evaluate the perfor-

mance of a model in terms of $L(h)$, as well as the following metrics:

1. *Risk Calibration*: A risk-calibrated model assigns risk predictions that match observed frequencies [126]. We measure risk calibration in terms of *expected calibration error*:

$$\text{ECE}(h) = \sum_{b=1}^{B} \frac{n_b}{n} |\hat{p}_b(h) - \bar{p}_b|. \tag{2.2}$$

   Here: I_b is the index set of n_b examples in bin $b \in [B]$; and $\hat{p}_b(h) := \frac{1}{n_b} \sum_{i \in I_b} h(\boldsymbol{x}_i)$ and $\bar{p}_b = \frac{1}{n_b} \sum_{i \in I_b} 1[y_i = +1]$ are the mean predicted risk and mean observed risk of examples in bin $b \in [B]$, respectively.

2. *Rank Accuracy*: A rank-accurate model outputs risk predictions that can be used to correctly order examples in terms of true risk. We assess rank accuracy using the *area under the ROC curve*:

$$\text{AUC}(h) = \frac{1}{n^+ n^-} \sum_{\substack{i:y_i=+1 \\ k:y_k=-1}} 1[h(\boldsymbol{x}_i) > h(\boldsymbol{x}_k)], \tag{2.3}$$

   where $n^+ = |\{i : y_i = +1\}|$ and $n^- = |\{i : y_i = -1\}|$.

In what follows, we let $M(h; \mathcal{D}) \in \mathbb{R}_+$ denote the performance of $h \in \mathcal{H}$ over a dataset $\mathcal{D}$ in regards to *performance metric* $M(g)$, where the convention is that lower values of $M(g)$ are better; when working with AUC, we measure the *AUC error*: $M(g) = 1 - \text{AUC}(g)$.

2.3.1 Competing Models

Competing models are classifiers with near-optimal performance compared to the baseline model. A *competing model* is any model $g \in \mathcal{H}$ whose performance is within ε of the baseline model h_0.

Definition 1 (ε-Level Set)**.** *Given a baseline model h_0, metric M, and error tolerance $\varepsilon > 0$, the* set of

competing models *(ε-level set) is the set:*

$$\mathcal{H}_\varepsilon(h_0) := \{h \in \mathcal{H} : M(h) \leq M(h_0) + \varepsilon\}.$$

Our methods consider multiplicity over a range of ε values. In practice, a suitable choice of ε should reflect the epistemic uncertainty in the performance of the baseline model. For instance, one could employ bootstrap re-sampling to measure the model uncertainty due to sample variation or consider worst-case uncertainty through generalization bounds.

2.3.2 Measuring Viable Risk Predictions

To examine how multiplicity affects predictions, we define a range of viable risk estimates that can be assigned by competing models.

Definition 2 (Viable Prediction Range)**.** *The* viable prediction range *is the smallest and largest risk estimate assigned to example i over competing models in the ε-level set:*

$$V_\varepsilon(\boldsymbol{x}_i) := [\min_{h \in \mathcal{H}_\varepsilon(h_0)} h(\boldsymbol{x}_i), \max_{h \in \mathcal{H}_\varepsilon(h_0)} h(\boldsymbol{x}_i)]. \tag{2.4}$$

For a prediction task, computing the viable prediction ranges over a sample illuminates the extent to which competing models assign different risk estimates to individuals. Although we express the prediction range over an ε-level set using $[\cdot, \cdot]$ interval notation, not all predictions between the min and the max may be attainable by a competing model.

2.3.3 Measuring Predictive Multiplicity

We say that a risk estimate is *conflicting* if it differs from the baseline risk estimate by at least some deviation threshold, $\delta \in (0, 1)$. The appropriate value of δ will depend on the application; i.e. a

conflicting risk prediction in a clinical decision support task may differ from that which constitutes a conflicting risk prediction in recidivism prediction. The following example illustrates the importance of reporting predictive multiplicity over a range of δ values.

Recidivism Prediction: Consider predicting an individual's risk of failing to appear in court using past arrest data [106]. Suppose there are four risk categories partitioned as follows– low: 0-10%, medium-low: 10-20%, medium-high: 20-30%, high: 30-100%. In this example, a deviation threshold $\delta = 10\%$ is informative because it would flag a change in risk that is large enough for any individual to go from "low" risk to "high" risk.

Medical Risk Prediction: Consider the task of predicting stroke risk for patients with atrial fibrillation (see e.g., the $CHADS_2$ risk score at MDCalc.com). The individual risk estimates can be used to inform blood thinner prescription decisions. One recommended usage suggests the following partitioning– 0% - 0.3%: do not prescribe blood thinner, 0.3-2.8%: maybe prescribe blood thinner, 2.9%+: prescribe blood thinner. If we study predictive multiplicity for this model, a value such as $\delta = 1\%$ is informative because a risk estimate shift by 1% could change the decision to prescribe a blood thinner for many individuals.

With a better understanding of the deviation threshold, we now define measures of predictive multiplicity. Ambiguity and discrepancy reflect the proportion of examples in a sample S assigned conflicting risk estimates by competing models. These definitions follow Marx et al. [112], who give analogous definitions for the problem of multiplicity with binary predictions (see Figure 2.1 for an illustration of the difference between this problem and the multiplicity of risk estimates).

Definition 3 (Ambiguity)**.** *The* (ε, δ)-ambiguity *of a probabilistic classification task over a sample* S *is the proportion of examples in* S *whose baseline risk estimate changes by at least* δ *over the* ε*-level set:*

$$\mathcal{A}_{\delta,\varepsilon}(h_0; S) := \frac{1}{|S|} \sum_{i \in S} 1\Big[\max_{h \in \mathcal{H}_\varepsilon(h_0)} |h(\boldsymbol{x}_i) - h_0(\boldsymbol{x}_i)| \geq \delta\Big].$$

Relative to the baseline model, ambiguity makes a statement about the proportion of individuals whose risk estimate is uncertain by at least δ. High ambiguity means more uncertainty in risk predictions. Users may also consult the viable prediction range to guide decisions using the baseline model.

Definition 4 (Discrepancy)**.** *The (ε, δ)-discrepancy of a probabilistic classification task over a sample S is the maximum proportion of examples in S whose risk estimates could change by at least δ by switching the baseline model with a competing model in the ε-level set:*

$$D_{\delta,\varepsilon}(h_0; S) := \max_{h \in \mathcal{H}_\varepsilon(h_0)} \frac{1}{|S|} \sum_{i \in S} 1[|h(\boldsymbol{x}_i) - h_0(\boldsymbol{x}_i)| \geq \delta].$$

Relative to the baseline model, discrepancy reflects the maximum the number of conflicting risk estimates as a result of replacing baseline model with a competing model in the ε-level set.

Ambiguity and discrepancy differ in the stance they take in regard to the worst case. Discrepancy measures the worst-case number of predictions that will change by switching the baseline model with a competing model. In contrast, ambiguity focuses on the worst case for prediction variation over the set of competing models. If we were to abstain from prediction on points that are assigned a conflicting prediction by a competing model using e.g., selective classification methods[21], then ambiguity would reflect the abstention rate.

Computing Ambiguity with Viable Prediction Ranges. As shown in Figure 2.2, we can use the viable prediction ranges of all points in a sample to compute ambiguity. Given the viable prediction range for each example, we can calculate the maximum difference between the baseline risk and that assigned by competing models. We can then compute ambiguity by measuring the proportion of examples where this difference exceeds the deviation threshold.

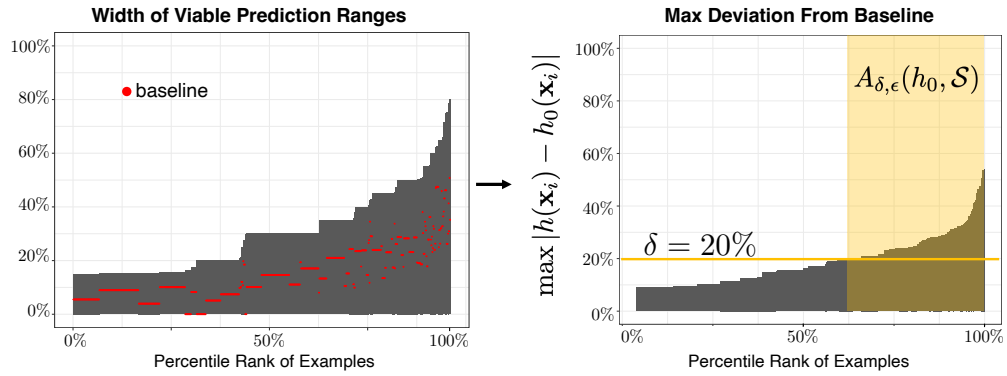


Figure 2.2: An illustration of how viable prediction ranges relate to ambiguity. Left, we plot the width of the viable prediction ranges $|V_\varepsilon(\boldsymbol{x}_i)|$ on the y-axis for each example on the x-axis. Note that widths shifted to start from zero and examples shown in increasing order. The plot also shows the individual baseline risk estimates for each example in red $h_0(\boldsymbol{x}_i) - \min_{h \in \mathcal{H}_\varepsilon(h_0)} h(\boldsymbol{x}_i)$ (shifted similarly). To interpret, the first example from the left has a width of $\approx 15\%$ with a baseline risk estimate on the lower side of the range. The last example has a width of $\approx 80\%$ with a baseline risk estimate closer to the higher side of the range. Using the viable prediction ranges $V_\varepsilon(\boldsymbol{x}_i)$ directly, we can extract the maximum difference from the baseline. On the right, we plot the maximum deviation from the baseline, $\max |h(\boldsymbol{x}_i) - h_0(\boldsymbol{x}_i)|$ for each example on the x-axis (increasing order). To interpret, consider a deviation threshold $\delta = 20\%$, all examples with max deviation above that threshold are highlighted in yellow giving us ambiguity, $A_{\delta,\varepsilon}(h_0; S)$.

2.4 Methodology

In this section, we detail the procedure for computing measures of predictive multiplicity. This methodology can be applied to any convex loss function $L(\cdot)$, and together with a training problem that employs a convex regularization term. We illustrate the methodology on the classification task described in §2.3 by training a probabilistic classifier via logistic regression, with $h(\boldsymbol{x}_i) = \frac{1}{1+\exp(-\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle)}$, where $\boldsymbol{w} = [w_0, w_1, \ldots, w_d]^\top \in \mathbb{R}^{d+1}$ is a coefficient vector. We train this baseline model by solving Eq. (2.1) to minimize normalized *logistic loss*: $L(\boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-\langle \boldsymbol{w}, y_i \boldsymbol{x}_i \rangle))$.

2.4.1 Measuring Ambiguity

We first present a method for computing ambiguity for different choices of ε and δ. The method also gives a conservative approximation of the viable prediction range for each example. We construct a pool of *candidate models* that assign a specific risk estimate to each example. From these models, we select those with performance within ε of the baseline model as the set of competing models.

Definition 5 (Candidate Model)**.** *Given a baseline model h_0, a finite set of user-specified threshold probabilities $P \subseteq [0,1]$, then for each $p \in P$ a* candidate model *for example $\boldsymbol{x}_i$ is an optimal solution to the following constrained ERM:*

$$\begin{aligned} \min_{\boldsymbol{w}\in\mathbb{R}^{d+1}} \; & L(\boldsymbol{w}) \\ \text{s.t.} \quad & h(\boldsymbol{x}_i) \leq p, \quad \textit{if } p < h_0(\boldsymbol{x}_i) \\ & h(\boldsymbol{x}_i) \geq p. \quad \textit{if } p > h_0(\boldsymbol{x}_i) \end{aligned} \tag{2.5}$$

For each threshold probability $p \in P$, we train a candidate model h such that the probability assigned to the example is constrained to the threshold p. In this way, by training for each example and threshold probability $p \in P$, we obtain the set of candidate models $\mathcal{G} := \{h : i \in S, p \in P\}$. We choose to solve the instances in order of increasing values of threshold probability p, which allows us to warm-start the optimization using previous solutions.

Given the set of candidate models, we define a *candidate ε-level set* as

$$\tilde{\mathcal{H}}_\varepsilon(h_0) = \{h \in \mathcal{G} : M(h) \leq M(h_0) + \varepsilon\}. \tag{2.6}$$

We use the candidate ε-level set to compute measures of predictive multiplicity. This method is exact for ambiguity defined in terms of near-optimal loss when the grid of threshold probabilities $P \subseteq [0,1]$ aligns with $h_0(x_i) \pm \delta$ (i.e., is selected as appropriate to the baseline prediction for

an example and the value of δ). For other metrics, such as AUC, this approach to compute ambiguity gives a conservative estimate (i.e., lower bound)—the training of a candidate model does not directly optimize for AUC, but we can retain only those candidate models that are competitive for the appropriate ε-level set definition. Since $\tilde{\mathcal{H}}_\varepsilon(h_0) \subseteq \mathcal{H}_\varepsilon(h_0)$, the candidate-model approach also provides a conservative estimate of the viable prediction range (Eq. (2.4)) for an example.

2.4.2 Measuring Discrepancy

Discrepancy is the maximum proportion of examples assigned conflicting risk estimates by a single competing model, $h \in \mathcal{H}_\varepsilon(h_0)$. Recall that a conflicting risk estimate differs from the baseline risk estimate $h_0(\boldsymbol{x}_i)$ by at least some deviation threshold, $\delta > 0$. Therefore, measuring discrepancy with respect to a baseline model corresponds to solving the following maximization problem:

$$\max_{h \in \mathcal{H}_\varepsilon(h_0)} \quad \sum_{i \in S} 1[|h(\boldsymbol{x}_i) - h_0(\boldsymbol{x}_i)| \geq \delta]. \tag{2.7}$$

Given a sample S, the baseline loss L_0, error tolerance ε, and deviation threshold δ, we can formulate Eq. (2.7) as a mixed-integer non-linear program (MINLP):

$$\begin{aligned}
\max_{\boldsymbol{w} \in \mathbb{R}^{d+1}} \quad & \sum_{i \in S} d_i & & \\
\text{s.t.} \quad & L(\boldsymbol{w}) \leq L_0 + \varepsilon & & \text{(2.8a)} \\
& d_i = v_{i,\delta} + z_{i,\delta} & \forall i \in S & \text{(2.8b)} \\
& M_{z,i}(1 - z_{i,\delta}) \geq \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - U_{i,\delta} & \forall i \in S & \text{(2.8c)} \\
& M_{v,i}(1 - v_{i,\delta}) \geq -\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + B_{i,\delta} & \forall i \in S & \text{(2.8d)} \\
& d_i, z_{i,\delta}, v_{i,\delta} \in \{0, 1\} & \forall i \in S &
\end{aligned}$$

The MINLP in (2.8) fits the parameters of a linear classifier that maximizes discrepancy . Here,

the objective maximizes number of examples assigned a conflicting risk estimate using the indicator variables $d_i := 1[|h(\boldsymbol{x}_i) - h_0(\boldsymbol{x}_i)| \geq \delta]$. Each d_i is set to $z_{i,\delta} := 1[h(\boldsymbol{x}_i) \leq (h_0(\boldsymbol{x}_i) - \delta)]$ (or $v_{i,\delta} := 1[h(\boldsymbol{x}_i) \geq (h_0(\boldsymbol{x}_i) + \delta)]$) when the model assigns a risk estimate to example i that exceeds δ on the low-side (or high-side) of the baseline risk estimate, respectively. We ensure the indicator behavior of $z_{i,\delta}$ and $v_{i,\delta}$ through the "Big-M" constraints (2.8d) and (2.8c), which flag deviations in score space. The Big-M parameters can be set as $M_{z,i} := -U_{i,\delta} + \max_{\boldsymbol{w}} \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle$ and $M_{v,i} := B_{i,\delta} - \min_{\boldsymbol{w}} \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle$, where $U_{i,\delta} := \text{logit}(h_0(\boldsymbol{x}_i) - \delta)$, and $B_{i,\delta} := \text{logit}(h_0(\boldsymbol{x}_i) + \delta)$. When the values of $U_{i,\delta}$ and $B_{i,\delta}$ lie outside of the $[0, 1]$ domain of the logit, we can drop the relevant indicator variable from the formulation. We provide additional details in the below.

MIP Formulation for Discrepancy

To train a competing model that optimizes discrepancy, we solve a maximization problem of the form:

$$\max_{h \in \mathcal{H}_\varepsilon(h_0)} \quad \sum_{i=1}^{n} d_i \tag{2.9}$$

Here, $d_i = 1[|h(\boldsymbol{x}_i) - h_0(\boldsymbol{x}_i)| \leq \delta]$ can also be rewritten in terms of score $d_i = 1[s_w(\boldsymbol{x}_i) \geq \text{logit}(\delta + h_0(\boldsymbol{x}_i))] + 1[s_w(\boldsymbol{x}_i) \leq \text{logit}(h_0(\boldsymbol{x}_i) - \delta)]$. We recover the solution to (2.9) by solving the following integer program:

$$\begin{aligned}
\max_{\boldsymbol{w} \in \mathbb{R}^{d+1}} \quad & \sum_{i=0}^{n} d_i \\
\text{s.t.} \quad & L(\boldsymbol{w}) \leq L(\boldsymbol{w}_0) + \varepsilon && \text{(2.10a)} \\
& d_i = v_{i,\delta} + z_{i,\delta} \qquad i = 1, \ldots, n && \text{(2.10b)}
\end{aligned}$$

$$M_{z,i}(1 - z_{i,\delta}) \geq (s_w(\boldsymbol{x}_i) - U_{i,\delta}) \qquad i = 1, ..., n \tag{2.10c}$$

$$M_{v,i}(1 - v_{i,\delta}) \geq -(s_w(\boldsymbol{x}_i) - B_{i,\delta}) \qquad i = 1, ..., n \tag{2.10d}$$

$$s_w(\boldsymbol{x}_i) = \sum_{j=0}^{d} w_j x_{i,j} \qquad i = 1, ..., n \tag{2.10e}$$

$$d_i \in \{0, 1\} \qquad i = 1, ..., n \tag{2.10f}$$

$$z_{i,\delta} \in \{0, 1\} \qquad i = 1, ..., n \tag{2.10g}$$

$$v_{i,\delta} \in \{0, 1\} \qquad i = 1, ..., n \tag{2.10h}$$

$$w_j \in \mathbb{R} \qquad j = 0, ..., d \tag{2.10i}$$

Here:

- $L(\boldsymbol{w}_0) := \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-\langle \boldsymbol{w}_0, y_i \boldsymbol{x}_i \rangle))$ is the log-loss of the baseline classifier on the training data
- $\varepsilon \geq 0$ is the loss tolerance (i.e., the maximum additional loss of any competing classifier)
- $U_{i,\delta}$ is a parameter that we set as $U_{i,\delta} := \text{logit}(h_0(\boldsymbol{x}_i) - \delta)$
- $B_{i,\delta}$ is a parameter that we set as $B_{i,\delta} := \text{logit}(h_0(\boldsymbol{x}_i) + \delta)$
- $M_{z,i}$ is a Big-M parameter that we set as $M_{z,i} = -U_{i,\delta} + \max_{\boldsymbol{w}} \sum_{j=0}^{d} w_j x_{ij}$
- $M_{v,i}$ is a Big-M parameter that we set as $M_{v,i} = B_{i,\delta} - \min_{\boldsymbol{w}} \sum_{j=0}^{d} w_j x_{ij}$
- $W^{\max}$ and $W^{\min}$ are user-defined coefficient bounds

Big-M Derivations Recall that by definition,

$$h(\boldsymbol{x}_i) := \Pr(y_i = +1 | \boldsymbol{x}_i) = \frac{1}{1 + \exp(-\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle)} \tag{2.11}$$

Therefore, $s_w(\boldsymbol{x}) = \text{logit}(h(\boldsymbol{x}_i))$. Our goal is to write the objective, $|h(\boldsymbol{x}_i) - h_0(\boldsymbol{x}_i)| \geq \delta$, in terms of score, $s_w(\boldsymbol{x}_i)$.

$$
\begin{aligned}
d_i &= 1[|h(\boldsymbol{x}_i) - h_0(\boldsymbol{x}_i)| \geq \delta], \\
&= 1[h(\boldsymbol{x}_i) - h_0(\boldsymbol{x}_i) \geq \delta] + 1[h_0(\boldsymbol{x}_i) - h(\boldsymbol{x}_i) \geq \delta] \\
&= 1[h(\boldsymbol{x}_i) \geq \delta + h_0(\boldsymbol{x}_i)] + 1[-h(\boldsymbol{x}_i) \geq \delta - h_0(\boldsymbol{x}_i)] \\
&= 1[h(\boldsymbol{x}_i) \geq h_0(\boldsymbol{x}_i) + \delta] + 1[h(\boldsymbol{x}_i) \leq h_0(\boldsymbol{x}_i) - \delta]
\end{aligned}
$$

Now we transform into score space

$$
\begin{aligned}
&= 1[\text{logit}(h(\boldsymbol{x}_i)) \geq \text{logit}(h_0(\boldsymbol{x}_i) + \delta)] + 1[\text{logit}(h(\boldsymbol{x}_i)) \leq \text{logit}(h_0(\boldsymbol{x}_i) - \delta)] \\
&= 1[s_w(\boldsymbol{x}_i) \geq \text{logit}(h_0(\boldsymbol{x}_i) + \delta)] + 1[s_w(\boldsymbol{x}_i) \leq \text{logit}(h_0(\boldsymbol{x}_i) - \delta)]
\end{aligned}
$$

Let $U_{i,\delta} = \text{logit}(h_0(\boldsymbol{x}_i) - \delta)$ and $B_{i,\delta} = \text{logit}(h_0(\boldsymbol{x}_i) + \delta)$.

$$
\begin{aligned}
&= 1[s_w(\boldsymbol{x}_i) \geq B_{i,\delta}] + 1[s_w(\boldsymbol{x}_i) \leq U_{i,\delta}] \\
&= v_{i,\delta} + z_{i,\delta}
\end{aligned}
$$

To ensure that $z_{i,\delta} = 1$ whenever $1[s_w(\boldsymbol{x}_i) \leq U_{i,\delta}] = 1$, and $z_{i,\delta} = 0$ whenever $1[s_w(\boldsymbol{x}_i) \leq U_{i,\delta}] = 0$, we add the following Big-M constraint:

$$
M_{z,i}(1 - z_{i,\delta}) \geq s_w(\boldsymbol{x}_i) - U_{i,\delta}
$$

Here we can set the Big-M parameter as:

$$\begin{aligned}
M_{z,i} &= \max_{\boldsymbol{w}}(s_w(\boldsymbol{x}_i) - U_{i,\emptyset}), \\
&= -U_{i,\emptyset} + \max_{\boldsymbol{w}} s_w(\boldsymbol{x}_i), \\
&= -U_{i,\emptyset} + \max_{\boldsymbol{w}} \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle, \\
&= -U_{i,\emptyset} + \max_{\boldsymbol{w}} \sum_{j=0}^{d} w_j x_{ij} \\
&= -U_{i,\emptyset} + W^{\max} \sum_{j=0}^{d} x_{ij}
\end{aligned}$$

Next, to ensure that $v_{i,\emptyset} = 1$ whenever $1[s_w(\boldsymbol{x}_i) \geq B_{i,\emptyset}] = 1$, and that $v_{i,\emptyset} = 0$ whenever $1[s_w(\boldsymbol{x}_i) \geq B_{i,\emptyset}] = 0$, we add the following Big-M constraint:

$$M_{v,i}(1 - v_{i,\emptyset}) \geq -(s_w(\boldsymbol{x}_i) - B_{i,\emptyset})$$

Here, we can set the Big-M parameter as:

$$\begin{aligned}
M_{v,i} &= \max_{\boldsymbol{w}}(B_{i,\emptyset} - s_w(\boldsymbol{x}_i)), \\
&= B_{i,\emptyset} + \max_{\boldsymbol{w}} -s_w(\boldsymbol{x}_i), \\
&= B_{i,\emptyset} - \min_{\boldsymbol{w}} s_w(\boldsymbol{x}_i), \\
&= B_{i,\emptyset} - \min_{\boldsymbol{w}} \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle, \\
&= B_{i,\emptyset} - \min_{\boldsymbol{w}} \sum_{j=0}^{d} w_j x_{ij}, \\
&= B_{i,\emptyset} - W^{\min} \sum_{j=0}^{d} x_{ij}
\end{aligned}$$

Outer-Approximation Algorithm

The challenge in solving (2.8) is that constraint (2.8a) is non-linear. We construct a linear approximation of the loss see e.g., [57,80] using an iterative, outer-approximation method see e.g., [163,13,12] to solve. The algorithm recovers a globally optimal solution to the MINLP in (2.8), and can be implemented using a mixed-integer programming solver with callback functions see e.g., [163,13,12]. The procedure builds a branch-and-bound tree to discover integer-feasible solutions that obey all constraints other than (2.8a). For each feasible solution identified, the procedure computes its loss to determine if it is feasible with respect to constraint (2.8a). If feasible, the procedure retains the solution. Otherwise, it updates the loss function approximation by adding a new linear constraint.

This method is exact for computing discrepancy in terms of near-optimal loss. For other metrics, we can again treat the intermediate solutions to the outer-approximation algorithm as candidate models and use these candidates to recover a lower bound similar to the method used in § 2.4.1.

Now, we provide the technical details of our outer approximation implementation.

Loss Callback Formulation We let $L_\varepsilon^{\max} := L^0 + \varepsilon$. This allows us to write the loss constraint $L(\boldsymbol{w}) \leq L^0 + \varepsilon$ as follows.

$$L(\boldsymbol{w}) \leq L_\varepsilon^{\max} \tag{2.12}$$

$$L(\boldsymbol{w}) - L_\varepsilon^{\max} \leq 0 \tag{2.13}$$

$$c(\boldsymbol{w}) \leq 0 \tag{2.14}$$

We will present an algorithm where we approximate $c(\cdot)$ by a linear approximation at a fixed

point $\boldsymbol{w}^k \in \mathbb{R}^d$. The linear approximation has the form:

$$\hat{c}^k(\boldsymbol{w}) := c(\boldsymbol{w}^k) + \nabla L(\boldsymbol{w}^k)(\boldsymbol{w} - \boldsymbol{w}^k) \tag{2.15}$$

$$= c(\boldsymbol{w}^k) + \sum_{j=1}^{d} \nabla L(w_j^k)(w_j - w_j^k) \tag{2.16}$$

Recall that $L(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(-\langle \boldsymbol{w}^k, y_i \boldsymbol{x}_i \rangle))$. The derivative evaluated at $\boldsymbol{w}^k$ is therefore,

$$\nabla_j L(w_j^k) = \frac{1}{n}\sum_{i=1}^{n} \nabla_j \log(1 + \exp(-\langle \boldsymbol{w}^k, y_i \boldsymbol{x}_i \rangle)) \tag{2.17}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \left[\frac{1}{1 + \exp(-\langle \boldsymbol{w}^k, y_i \boldsymbol{x}_i \rangle)} * \exp(-\langle \boldsymbol{w}^k, y_i \boldsymbol{x}_i \rangle) * -y_i \boldsymbol{x}_i \right] \tag{2.18}$$

To perform the outer approximation, we add the following loss cut if $L(\boldsymbol{w}^k) - L_\varepsilon^{\max} > 0$

$$0 \geq L(\boldsymbol{w}^k) - L_\varepsilon^{\max} + \sum_{j=1}^{d} \nabla L(w_j^k)(w_j - w_j^k) \tag{2.19}$$

$$0 \geq L(\boldsymbol{w}^k) - L_\varepsilon^{\max} + \sum_{j=1}^{d} \nabla L(w_j^k) * w_j - \sum_{j=1}^{d} \nabla L(w_j^k) * w_j^k \tag{2.20}$$

$$-\sum_{j=1}^{d} \nabla L(w_j^k) * w_j \geq L(\boldsymbol{w}^k) - L_\varepsilon^{\max} - \sum_{j=1}^{d} \nabla L(w_j^k) * w_j^k \tag{2.21}$$

2.5 Numerical Experiments

In this section, we present experiments on synthetic and real-world data. Our goals are to: (1) reveal dataset characteristics that impact predictive multiplicity; and (2) determine the extent to which real risk assessment tasks exhibit predictive multiplicity in practice.

Name	Outcome Variable	n	d	Class Imbalance	Train Loss	Train AUC	Train ECE
mammo[52]	mammogram shows breast cancer	961	12	0.86	0.471	85%	2.4%
credit[182]	customer default on loan	30,000	23	3.50	0.453	74%	1.6%
bank[123]	person opens bank account after marketing call	41,188	57	0.12	0.268	82%	0.9%
adult[93]	person in 1994 US census earns over $50,000	32,561	36	0.31	0.332	90%	0.8%
compas_arrest[3]	rearrest for any crime	5,380	18	0.84	0.612	72%	1.1%
compas_violent[3]	rearrest for violent crime	8,768	18	0.13	0.332	67%	0.3%
apnea[165]	patient diagnosed with obstructive sleep apnea	1,537	36	0.70	0.565	76%	3.3%

Table 2.1: Publicly available datasets used to train risk assessment models in §2.5.2. For each dataset, we report n, d, the class imbalance ratio, $|n^+|/|n^-|$, and the performance metrics of the baseline model on training data. We work with sub-sampled versions of credit, bank and adult by randomly sampling $n = 5000$ points from each dataset.

2.5.1 Synthetic Datasets

Linear Separability. To demonstrate how separability informs predictive multiplicity, we compute ambiguity while varying the degree of separability and show results in Figure 2.3 column **(A)**. We set $\delta = 20\%$ and $\varepsilon = 5\%$ and control separability by increasing the variance of the data from $\sigma = 4$ (top) to $\sigma = 10$ (bottom). A clear trend is that ambiguity increases as the data becomes less separable from 1% to 21%. Notice, also that the ambiguous examples tend to be those near the discriminant boundary and outliers.

Outliers and Margin Distance. We examine how predictive multiplicity relates to outlier distance from the discriminant boundary. We position outliers near and far from the discriminant boundary and compute ambiguity. As shown in Figure 2.3 column **(B)**, a clear trend is that examples that are outliers but far from the discriminant boundary (high margin) are less susceptible to predictive multiplicity.

Majority-Minority Structure. We consider the effect of systematically varying the majority-minority structure of data. For this, we generate a majority class that has a different statistical pattern of features than a minority class. Given the two groups, the model is faced with a tradeoff between correctly predicting one group or the other. In Figure 2.3 column **(C)**, we vary the ratio in a majority-minority structure revealing that the minority group is more prone to predictive multiplicity. The ambiguity of the minority group at 10:1 is substantially larger than for the majority group. This shows the importance of evaluating multiplicity across subgroups.

2.5.2 Real-World Datasets

In this section, we evaluate predictive multiplicity in risk prediction tasks from medicine, lending, and criminal justice.* Altogether, we consider seven datasets that exhibit variations in sample size, number of features, and class imbalance. For each dataset, we compute viable prediction ranges, ambiguity and discrepancy using the methods outlined in §2.4. When training candidate models, we adopt a grid of target predictions: $P = \{0.01, 0.1, 0.2, \ldots, 0.9, 0.99\}$. We compute discrepancy by solving the MINLP Eq. (2.7) with CPLEX v20.1 [46] on a single CPU with 16GB RAM. Our results are shown in Figure 2.4.

*This is not an endorsement of current usage of risk assessment tools in criminal justice. The use of prediction software raises serious concerns in this domain. We do not condone building models on arrest data to inform or justify increased policing.

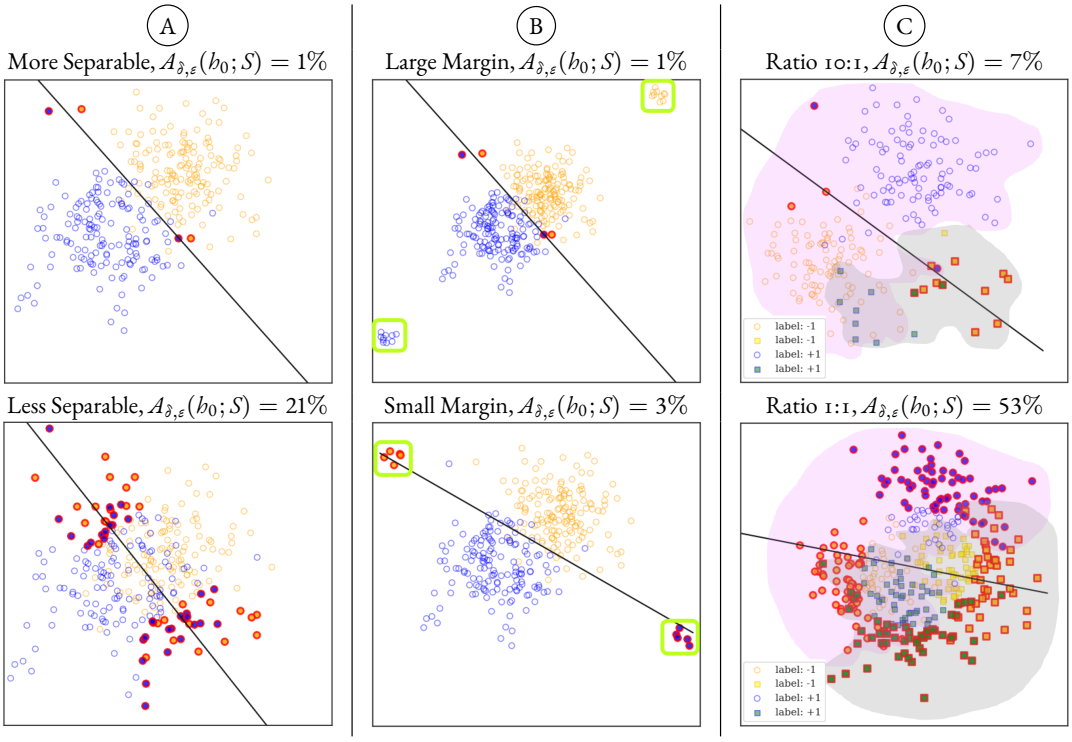


Figure 2.3: Experiments on synthetic data. In (A), we vary separability and find that ambiguity increases as separability decreases. In (B), we position outliers near and away from the discriminant boundary finding that outliers closer to the boundary are more prone to ambiguity. In (C), we vary the ratio in a majority-minority structured dataset: magenta shading- majority group (circles), grey shading- minority group (squares) revealing that the minority group is more prone to ambiguity. In the figures, $Y = +1$ examples are blue, $Y = -1$ examples are orange, and ambiguous examples are highlighted red and we set $\partial = 20\%$ and $\varepsilon = 5\%$.

Viable Prediction Ranges. Our results show that competing models can assign risk estimates that vary substantially. Viable prediction ranges are plotted in rows **(A)** and **(B)** of Figure 2.4, and we see non-zero viable prediction ranges for all examples across all datasets. The viable ranges for apnea and mammo appear much larger compared to compas_arrest. In terms of near-optimal loss, apnea has the most variation, while mammo has the most variation in terms of AUC. This points to the value in varying near-optimal metric.

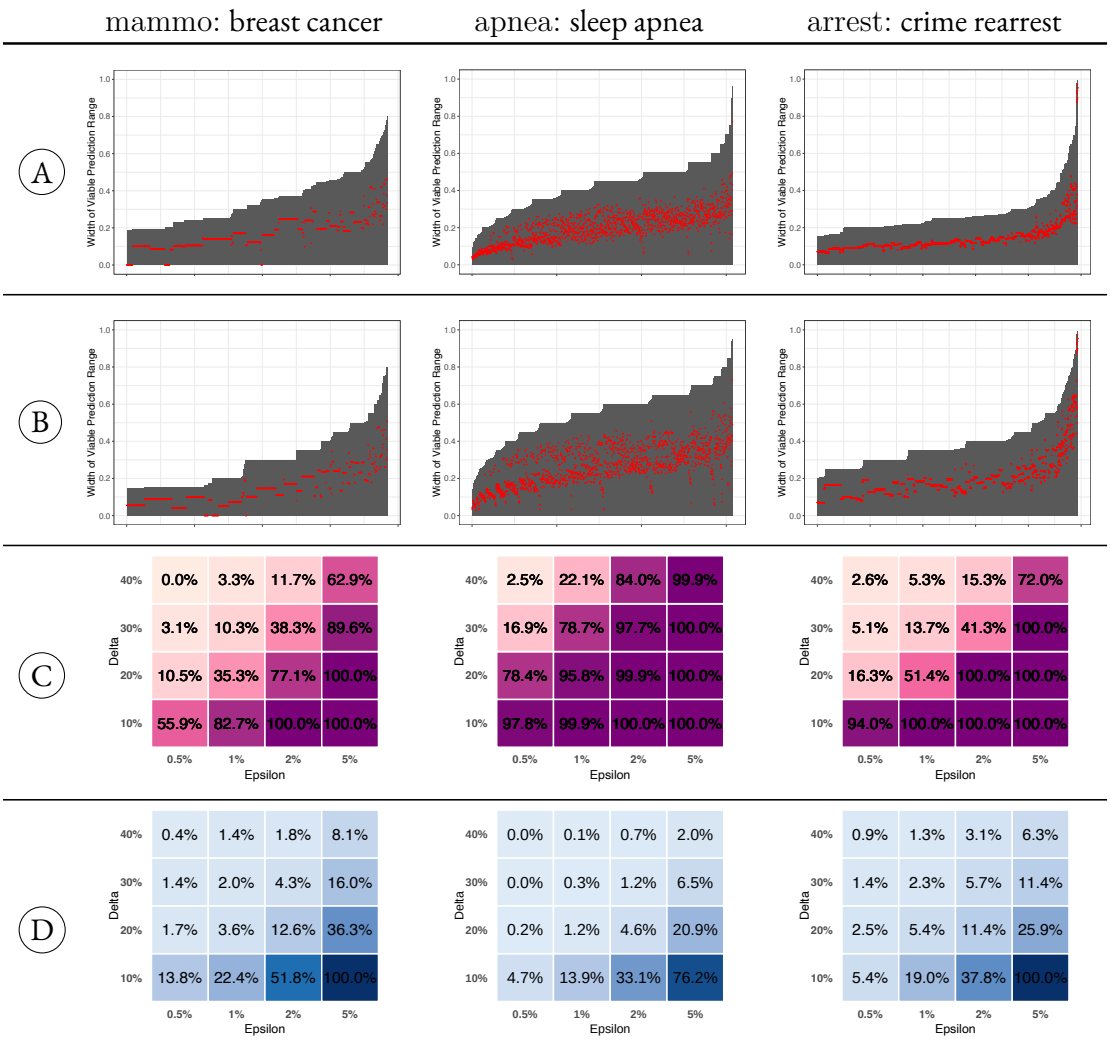


Figure 2.4: Predictive multiplicity in probabilistic classification on mammo, apnea and arrest. In rows (A) and (B) we show the distribution of viable prediction ranges $|V_\varepsilon(\boldsymbol{x}_i)|$ on the y-axis for each example on the x-axis (relative baseline estimates in red). Notice, pointwise viable prediction ranges are plotted in increasing order from left to right. We plot viable prediction ranges for competing models with near-optimal training AUC (A) and training loss (B). See illustration in Figure 2.2. We also show ambiguity (C) and discrepancy (D) for competing models with respect to training loss.

AMBIGUITY AND DISCREPANCY. Ambiguity and discrepancy are shown in rows **(C)** and **(D)** of Figure 2.4, respectively. For $\varepsilon = 1\%$ and $\delta = 20\%$, we see ambiguity values at 35.3% (mammo), 95.8% (apnea), and 51.4% (compas_arrest). This means that 35.3% of breast cancer risk estimates

vary by at least 20% over near-optimal models. We see discrepancy values at 3.6% (mammo), 1.2% (apnea), and 5.4% (compas_arrest) for $\varepsilon = 1\%$ and $\delta = 20\%$. compas_arrest is the worst in terms of discrepancy, while apnea has the most severe ambiguity. Thus, ambiguity and discrepancy are not always coupled.

On the Choice of Performance Metric. In settings where we want a model that performs well in terms of AUC, we should measure predictive multiplicity over a set of competing models with near-optimal AUC. In practice, it is often convenient to measure predictive multiplicity over a set of competing models that attain near-optimal loss (since the loss can be encoded into an optimization problem). This is a problem because small variations in loss can lead to large variations in AUC – thus models with near-optimal loss may not match models with near-optimal AUC. Our results show that measures of predictive multiplicity vary considerably based on the performance metric used to define the set of competing models. In particular, we find that discrepancy and ambiguity will vary when measured over competing models that attain near-optimal loss, AUC, or ECE.

For example, if we want to estimate the prevalence of samples whose predictions can change by over $\delta = 20\%$ on the mammo dataset, we find that ambiguity $= 35\%$ for competing models with loss within 1% of the baseline loss, but ambiguity $= 45\%$ over models with AUC within 0.5% of the baseline AUC. These differences highlight the need for approaches that measure predictive multiplicity in the terms of performance metric that we use to evaluate the model (e.g., AUC or ECE).

On Samples Prone to Ambiguity. Our results reveal a relationship between ambiguity and individual *uniqueness* (number of duplicates), *class* imbalance, and *baseline risk estimate*. Table 2.2 compares proportion of ambiguous examples in different subgroups. The subgroups are as follows: label = +1, label = -1, zero duplicates, more than one duplicate, more than 20 duplicates, baseline as-

Loss ε-level Set Subgroup	ε: 0.005	ε: 0.01	ε: 0.02	ε: 0.05
Dataset Ambiguity	0.11	0.35	0.77	1.00
Y = +1	0.09	0.39	0.78	1.00
Y = -1	0.12	0.32	0.76	1.00
Duplicates = 0	0.71	0.92	1.00	1.00
Duplicates > 1	0.07	0.32	0.76	1.00
Duplicates > 20	0.00	0.08	0.63	1.00
Baseline prob < 10%	0.02	0.06	0.54	1.00
Baseline prob > 90%	0.00	0.00	1.00	1.00
Baseline prob [45% : 55%]	0.28	1.00	1.00	1.00

(a) Loss ε-level set: Data set mammo

AUC ε-level Set Subgroup	ε: 0.2%	ε: 0.5%	ε: 1%	ε: 2%
Dataset Ambiguity	0.19	0.46	0.86	1.00
Y = +1	0.15	0.40	0.86	1.00
Y = -1	0.22	0.50	0.86	1.00
Duplicates = 0	0.58	0.92	1.00	1.00
Duplicates > 1	0.17	0.43	0.86	1.00
Duplicates > 20	0.05	0.31	0.78	1.00
Baseline prob < 10%	0.00	0.06	1.00	1.00
Baseline prob > 90%	0.00	0.00	0.00	1.00
Baseline prob [45% : 55%]	0.42	1.00	1.00	1.00

(b) AUC ε-level set: Data set mammo

Loss ε-level Set Subgroup	ε: 0.005	ε: 0.01	ε: 0.02	ε: 0.05
Dataset Ambiguity	0.49	0.72	0.89	1.00
Y = +1	0.78	0.98	1.00	1.00
Y = -1	0.40	0.63	0.85	1.00
Duplicates = 0	0.66	0.82	0.95	1.00
Duplicates > 1	0.38	0.65	0.85	1.00
Duplicates > 20	0.00	0.50	0.70	1.00
Baseline prob < 10%	0.16	0.43	0.77	1.00
Baseline prob > 90%	0.41	0.91	1.00	1.00
Baseline prob [45% : 55%]	1.00	1.00	1.00	1.00

(c) Loss ε-level set: Data set adult

AUC ε-level Set Subgroup	ε: 0.2%	ε: 0.5%	ε: 1%	ε: 2%
Dataset Ambiguity	0.37	0.63	0.81	0.94
Y = +1	0.64	0.95	0.99	1.00
Y = -1	0.28	0.53	0.76	0.92
Duplicates = 0	0.55	0.76	0.91	0.99
Duplicates > 1	0.25	0.55	0.76	0.91
Duplicates > 20	0.00	0.37	0.57	0.76
Baseline prob < 10%	0.07	0.28	0.63	0.88
Baseline prob > 90%	0.20	0.87	1.00	1.00
Baseline prob [45% : 55%]	0.88	1.00	1.00	1.00

(d) AUC ε-level set: Data set adult

Loss ε-level Set Subgroup	ε: 0.005	ε: 0.01	ε: 0.02	ε: 0.05
Dataset Ambiguity	0.66	0.97	1.00	1.00
Y = +1	0.92	1.00	1.00	1.00
Y = -1	0.63	0.97	1.00	1.00
Duplicates = 0	0.68	0.97	1.00	1.00
Duplicates > 1	0.45	0.95	1.00	1.00
Duplicates > 20	nan	nan	nan	nan
Baseline prob < 10%	0.52	0.96	1.00	1.00
Baseline prob > 90%	1.00	1.00	1.00	1.00
Baseline prob [45% : 55%]	1.00	1.00	1.00	1.00

(e) Loss ε-level set: Data set bank

AUC ε-level Set Subgroup	ε: 0.2%	ε: 0.5%	ε: 1%	ε: 2%
Dataset Ambiguity	0.15	0.31	0.62	0.92
Y = +1	0.48	0.70	0.91	0.99
Y = -1	0.10	0.26	0.58	0.91
Duplicates = 0	0.15	0.33	0.64	0.93
Duplicates > 1	0.04	0.10	0.38	0.81
Duplicates > 20	nan	nan	nan	nan
Baseline prob < 10%	0.01	0.12	0.46	0.88
Baseline prob > 90%	0.00	1.00	1.00	1.00
Baseline prob [45% : 55%]	1.00	1.00	1.00	1.00

(f) AUC ε-level set: Data set bank

Table 2.2: The proportion of ambiguous points in different subgroup categories for data sets mammo, adult and bank, considering metrics of loss and AUC and for additive ε-level sets. The discrepancy threshold is set to $\Delta = 20\%$.

signed risk less than 10%, baseline assigned risk greater than 90%, and baseline assigned risk between 45% and 55%.

For uniqueness, we find that across datasets, less than 10% of examples with more than 20 duplicates are ambiguous. That unique examples are more prone to ambiguity is related to our findings on outliers (see §2.5.1).

In terms of class imbalance, we find datasets with class imbalance skewed negative (adult, bank) often exhibit multiplicity on positive examples. In comparison, datasets that are roughly balanced by class (e.g., mammo, compas_arrest) have the same level of ambiguity for each class. This can be interpreted in light of the majority-minority effect from §2.5.1.

In terms of the baseline risk estimate, we see high ambiguity for examples with baseline risk near 50% on all datasets. For instance, all examples with baseline risk between 45% and 55% are ambiguous for the mammo dataset ($\varepsilon = 0.5\%$ AUC, $\delta = 20\%$). There is no reason to believe that high ambiguity is less problematic for these samples. Rather, the importance of ambiguity will depend on the risk thresholds that drive decisions in a particular domain.

On the Disparate Impact of Multiplicity. Our results demonstrate how multiplicity can disproportionately impact individuals from historically marginalized groups. For example, when predicting the risk of rearrest, individuals who are ethnically Hispanic are disproportionately affected by predictive multiplicity: ambiguity is 39% for African Americans and 49% for Caucasians, compared to 98% for Hispanics ($\varepsilon = 1\%$ and $\delta = 20\%$). Hence, reporting predictive multiplicity for subgroups can reveal important fairness considerations when testing models deployed throughout society.

2.6 Concluding Remarks

We developed methods to evaluate the effect of slightly perturbing optimal model performance, revealing that similar models do not always assign similar predictions. We studied how competing models can assign conflicting predictions in probabilistic classification tasks. The proposed optimization-based methods compute our simple measures reliably. Compared to previous work, our methods allow for flexibility in choosing near-optimal metric and deviation threshold. Using synthetic data, we also present the first study providing insight into the kinds of data characteristics that give rise to predictive multiplicity and show that separability, outliers and majority-minority structure are informative. Empirically, we reveal concerning levels of predictive multiplicity in high-stakes domains.

More research is needed to examine predictive multiplicity for other loss functions and model classes (our methods immediately generalize to linear models with convex loss functions). Also, it will be important to study how to effectively communicate these effects to practitioners and decision makers. Also, when a practitioner encounters high predictive multiplicity, more work is needed on response options and mitigation strategies. Given predictive multiplicity metrics, practitioners can make better decisions in model selection while end-users can adjust their reliance on individual risk predictions. Concisely, analyzing predictive multiplicity promotes accountability and transparency in machine learning.

Asking the proper question is the central action of transformation- in fairy tales, in analysis, and in individuation. The key question causes germination of consciousness. The properly shaped question always emanates from an essential curiosity about what stands behind. Questions are the keys that cause the secret doors of the psyche to swing open.

Clarissa Pinkola Estés

3

Multi-Target Multiplicity: Flexibility and Fairness in Target Specification under Resource Constraints

3.1 Introduction

Scholars have argued that prediction problems are ubiquitous across many domains of decision-making, from employment, to education, to health[89]. Yet real-world problems rarely present themselves as fully formed machine learning tasks[139]. Critically, it is often not clear what target should be predicted to help decision makers achieve their goals[70,133]. For example, while it might seem self-evident that creditors should be predicting default, what constitutes "default" is not a given. Creditors need to make an affirmative choice about the number of months of missed payments that ultimately count as "default"[71]. In some cases, the decision is not based on just one chosen target, but instead a combination of targets. For example, many algorithmic tools currently used in criminal justice and human services function by aggregating predictions of several different targets, ranging from different types of criminal justice system encounters, to mental and physical health outcomes, to measures of housing stability[88,169].

A recent line of work has explored the implications of this flexibility in target variable choice for fairness. In particular, researchers have pointed out that different choices for the prediction target can lead to more or less disparity in selection rates across groups[133,129,77,121,125,111,87,55]. As dis-

cussed in the introduction of this thesis, one particularly well-known study by Obermeyer et al. [129] illustrates both the risks and benefits of target choice.

The existing literature offers several examples that highlight the potential importance of target variable choice. Prior work does not, however, offer a more general mathematical or computational framework for characterizing the extent to which target variable choice affects individuals' outcomes and selection rate disparities across groups. As discussed in the introduction, this chapter aims to fill this gap.

Prior work on multiplicity has at times explicitly steered clear of viewing target choice as a source of multiplicity, while at the same time acknowledging that it plays an important role in problem formulation more generally [22]. One reason for this is that multiplicity has been studied with respect to the task of predicting a pre-specified target. From this perspective, considering different outcomes amounts to considering a different task. In our setting, however, we adopt a broader view of the task for which multiplicity is being assessed. Specifically, we note that all the motivating examples we have just discussed can be thought of as *predictive allocation tasks*—tasks where historical data is used to learn a "prioritization" or "risk" score and where that score then serves as the basis for deciding how to allocate resources, usually as part of human-in-the-loop decision processes. In practice, predictive allocation tasks are governed by resource constraints. There is only a finite amount of benefit, burden, or scrutiny that the system is able to allocate. For instance, the algorithm investigated by Obermeyer et al. was developed to help allocate coordinated care management to a certain number of clients. Similarly, employers cannot offer jobs to *everyone* they predict will be a sufficiently good employee, whatever target or set of targets they choose to predict to make such an assessment. If a range of models trained to predict different targets can be similarly helpful in performing a predictive allocation task, then it is reasonable to understand the flexibility afforded by target choice in terms of multiplicity as well.

In addition to introducing a framework for multi-target multiplicity, we also refine the standard

treatment of predictive multiplicity in the single-target setting to account for additional practical constraints inherent in predictive allocation tasks. Given their limited budgets, they are likely only able to offer jobs to a select few applicants. This means that the set of "good models", whether in the single-target or multi-target setting, can only include models that satisfy the resource constraint. In this work we demonstrate how to introduce resource constraints into the study of multiplicity.

In summary, our work introduces the concept of multi-target multiplicity, and provides a formal and computational framework for quantifying the level of multiplicity that exists in a given predictive allocation task. Along the way we introduce a refinement of single-target predictive multiplicity to the resource constrained setting, and introduce corresponding computational methods. Our primary contributions are as follows.

1. We introduce a framework for assessing single-target multiplicity in the presence of resource constraints (§3.3). We define a new measure of predictive multiplicity (top-κ ambiguity) and present a mixed integer program (MIP) to calculate this ambiguity measure for linear models.

2. We introduce the concept of multi-target multiplicity alongside a framework for assessing multi-target multiplicity for predictive allocation tasks (§3.4.1).

3. We demonstrate how the framework can be used to assess fairness-related measures by presenting a MIP that calculates the minimum and maximum attainable selection rate for a given group (§3.4.4).

4. We demonstrate our framework on the healthcare dataset released by Obermeyer et al. and provide semi-synthetic experiments that aim to clarify how we might be able to improve fairness by moving to a multi-target setting (§3.6). We reiterate the original results showing that modeling active chronic conditions produces the highest concentration of current illnesses in the high-risk set. Given a comparatively small variation in outcome concentration across different target

variable choices, we do see a substantial difference in the racial composition of the high-risk set (more than 10 percent difference). Findings show that the index model captures each of the individual targets that it is comprised of and also produces a high-risk set with a high concentration of Black patients, as per the objective of the multi-target group selection formulation. Using the proposed framework, we arrive at a way of ranking patients that results in a more equitable allocation of a scarce resource via the index model.

3.2 Related Work

Problem formulation and fairness. Prior work has grappled with many aspects of problem formulation that have implications for fairness. Some scholars have focused on the fact that the underlying goals driving the process of developing a machine learning model can be normatively suspect, regardless of any particular properties of the resulting model [59,73,84]. Scholars have identified various reasons why the choice of target might raise concerns with fairness: some outcomes or qualities of interests might just be more evenly distributed across the population than others [133,87]; certain outcomes or qualities of interests might be easier to predict with similar degrees of accuracy across the population than others [39]; some kinds of selection bias might cause certain outcomes or qualities of interest to be observed more or less frequently in certain groups rather than others, even if they occur at similar rates in reality [107]; certain targets might suffer from more so-called "label bias" than others—that is, systematically less accurate observations of the true value of the target for members of some groups than others [78,40,77]. Indeed, one way to understand the Obermeyer et al. study is as a form of label bias since healthcare costs acted as a systematically inaccurate measure of underlying healthcare needs. Our work departs from much of this literature by focusing on cases where there is no obviously right or clearly preferable choice of target or proxy and thus uncertainty about which to choose or whether to choose more than one.

Multi-task learning, multi-criteria decision-making, latent variable modeling, and fairness. While our use of the term "multi-target" might suggest a close connection to fairness considerations in multi-task learning (see, e.g., [171]), the problem we study is distinct. Whereas in multi-task learning the goal is to perform well on (and assess fairness for) *each* of K prediction tasks by borrowing strength across tasks, in our setting we are interested in arriving at a *single model*, which may not perform optimally on any individual task, but which successfully captures multiple desiderata. In this sense, our setting is more closely related to recent work on latent variable modeling in recommender systems that aims to optimize for a latent *value* using a combination of noisy observed measures, such as clicks, replies, reshares, and other observable forms of user engagement [120,90]. A key difference is that we do not posit a specific notion of optimality, and instead explore the degree of multiplicity inherent in a class of learning procedures for forming a univariate prediction from multiple available targets. Lastly, our work connects to the extensive literature on multi-criteria decision-making (MCDM) in operations research. Indeed, the index model and index variable approaches we introduce in §3.4.1 parallel the classic *weighted sums* method of combining multiple criteria (e.g., loss or other objective functions) into a single objective [61]. However, whereas the focus of MCDM is in the values of the different objective functions, we examine multiplicity, which pertains to the variability in prediction decisions for *individual people or cases*. Additionally, a similar concern for arbitrariness that comes into machine learning through the choice of target arises in the context of university and college ranking where the ranking of a school depends on the choice of target by the agency responsible for ranking [65].

Predictive multiplicity and fairness. There is also a growing literature that seeks to explore the normative implications of multiplicity. Scholars have investigated the degree to which multiplicity can be leveraged to improve interpretability [147] and explainability [56,47,134]. Others have examined the danger multiplicity poses for robustness [51] and non-arbitrariness [19,22,144,76,36]. Still others have focused on its implications for fairness [112,1,40,22,176]. Notably, some of this work has defined mea-

sures and developed methods for evaluating predictive multiplicity in binary classification [112] and probabilistic classification [176,76], focusing on so called "ambiguity" in models' predictions (i.e., the amount of disagreement in models' predictions on different points). Our work is the first to extend the analysis of multiplicity to the problem of predictive allocation under resource constraints. We introduce measures of multiplicity for both single-target and multi-target settings, and introduce efficient methods that, for a subset of points, can certify whether those points contribute to the multiplicity measure.

We note that the resource allocation problem formulated in this chapter could, in theory, be reduced to a binary classifier that predicts whether an example is above or below a threshold. Analyzing predictive multiplicity in binary classification [112] has involved minimizing 0-1 loss directly. For instance, Marx et al. [112] compute ambiguity by training candidate models that minimize 0-1 loss such that a given prediction conflicts with the baseline prediction. Similar to Chapter 3, they then select from those candidate models those with near-optimal performance to form the Rashomon set. The integer programs formulated in Marx et al. [112] assume the binary classifier is logistic regression. In this chapter, if one thinks of our problem as analyzing a selection classifier (predicting top-κ selection decision), this classifier would be distinct. Here, the rank of individual examples is a function of a probabilistic classification output and predictive multiplicity metrics are defined in terms of this individual rank. Additionally, we incorporate the near-optimal performance constraint directly into the calculation of ambiguity and also analytically compute un-flippable examples removing them from consideration before computation whereas the formulation in Marx et al. [112] requires running the integer program for each example in the dataset without consideration for a resource constraint.

Resource constraints and fairness. Recent work on algorithmic fairness has noted the importance of considering resource constraints. For instance, Black et al. [18] discuss how the increased cost of auditing more complex tax filings can lead to prediction-based auditing strategies that disproportionately focus on lower income earners. Other work has emphasized the importance of considering

resource constraints in the context of algorithmic fairness in healthcare [137] and business analytics [43]. Our work provides a conceptual and computational framework for reasoning about fairness in the presence of resource constraints.

Group fairness. Our focus on selection rate disparities differs from traditional group fairness measures based on False Positive Rate (FPR) and True Positive Rate (TPR). FPR quantifies the proportion of negative examples wrongly classified as positive [54]. TPR is the proportion of positive examples correctly classified [54]. These measures are typically used when there is a focus on accuracy or error in classification. And in terms of group fairness, these measures can help gain insight into model performance across different subgroups. In the resource allocation setting, these measures would be somewhat indirect in that they characterize performance but not necessarily impact on resource distribution. The indirectness refers to the fact that these measures do not account for resource allocation constraints. Another important metric, consider demographic parity [50] or statistical parity that requires the rate of positive outcomes to be the same across subgroups (stratified by protected attribute). Satisfying demographic parity means each group would have an equal chance at receiving a positive classification. This focus on equality in rates of positive outcomes across subgroups does not consider prioritization in terms of whether each group has an equal chance at being selected to receive the resource. Simply because resource budget is not the direct focus. Selection rate disparity, which we focus on here, can be viewed as a variation in demographic parity that adjusts for only a limited number of instances being selected. Consider another important metric, equal opportunity [72], which requires that TPR be equal across subgroups. Equal opportunity specifically focuses on accuracy of positive outcomes instead of the distribution of the outcomes with respect to selection. In theory, it might be possible to achieve equal TPR while having unequal overall selection rates. For this reason, we opt to use selection rate directly to emphasize access to constrained resources instead of opting for traditional group fairness metrics.

3.3 Predictive multiplicity with resource constraints

In this section, we introduce a framework for examining single-target predictive multiplicity under resource constraints. Our goal is to study predictive consistency over models with near-optimal performance for each target option. We provide key definitions for predictive multiplicity in §3.3.2, present a computational framework based on mixed-integer programming (MIP) for linear models in §3.3.4 and introduce methods for improving computational efficiency in §3.3.5.

3.3.1 Preliminaries

We consider a dataset, $\mathcal{D} = \{(x_i, a_i, \tilde{y}_i^{(k)})\}_{i=1}^n$, consisting of n cases, where $x_i = [1, x_{i1}, \ldots, x_{id}] \in \mathcal{X} \subseteq \mathbb{R}^{d+1}$ is a feature vector, $y_i \in \mathbb{R}$ is an outcome of interest (potentially binary), and $a_i \in A$ is a protected attribute. We operate within the prediction-based allocation setting where a limited resource is to be allocated to instances in descending order of the predicted value $\hat{y}_i = \hat{y}(x_i)$. If case i is selected, it is allocated r_i resources. Let κ denote the resource cap, and let $i_{(j)} = i_{(j)}(\hat{y})$ denote the instance with the jth largest value of $\hat{y}_i$ (so that $i_{(1)}$ is the index with the largest predicted value). Let $\tau_i = \tau_i(\hat{y})$ denote the rank of instance i in *descending* order.

We assume that resources get allocated to instances $i_{(1)}, \ldots, i_{(J)}$, where J is the largest value such that $\sum_{j=1}^{J} r_{i_{(j)}} \leq \kappa$. The most common prediction-based allocation setting in practice is where there is simply a limit to the number of cases that can be selected (i.e., $r_i = 1 \, \forall i$, in which case $J = \kappa$). While we restrict our attention to this setting, all metrics and computational methods can be extended to general $r_i \in \mathbb{R}_{>0}$.

3.3.2 Measuring predictive multiplicity under resource constraints

Predictive multiplicity is the extent to which models with near-equivalent performance produce different predictions or classifications. The set of models under consideration is often referred to as

the set of "good models"[47].

In prior work, Marx et al.[112] introduced predictive multiplicity metrics for binary classification, and Watson-Daniels et al.[176] considered the setting of probabilistic classification. As in the standard predictive multiplicity setting[112], we begin with a *baseline model*, h_0, that is the solution to an empirical risk minimization (ERM) problem of the form $\min_{h \in \mathcal{H}} L(h; \mathcal{D})$, over a hypothesis class, $\mathcal{H}$, with loss $L(\ \cdot\ ; \mathcal{D})$. In this context, one can consider the ε-Rashomon set, which is the set of all models that achieve near-optimal loss.

Definition 6 (ε-Rashomon set)**.** *For a baseline model h_0 and error tolerance $\varepsilon > 0$, the ε-Rashomon set of competing models is:*

$$\mathcal{H}_\varepsilon(h_0) := \{h \in \mathcal{H} : L(h; \mathcal{D}) \leq L(h_0; \mathcal{D}) + \varepsilon\}.$$

In[112], $\mathcal{H}$ is assumed to be a class of binary classifiers, and one of the predictive multiplicity measures the authors introduce is the *ambiguity* of a prediction problem,

$$\alpha_\varepsilon(h_0) = \frac{1}{n} \sum_{i=1}^{n} \max_{h \in \mathcal{H}_\varepsilon(h_0)} 1[h(x_i) \neq h_0(x_i)].$$

Note that under this definition, a prediction problem will have high ambiguity if the positive classification rate, $\frac{1}{n}|\{i : h(x_i) = 1\}|$, differs greatly between h_0 and models in $\mathcal{H}_\varepsilon(h_0)$. That is, a high ambiguity may simply result from models that allocate a very different number of resources.

To define an analogous measure for the resource constrained setting, we need to compare models at the same resource cap, κ. Recall that, unlike in[112], we consider $\mathcal{H}$ that is a class of prediction models returning continuous values in $\mathbb{R}$, not binary classifiers. Given a prediction model h and resource

cap κ, let

$$Top_{(i,h,\kappa)} = 1[\tau_i(h) \leq \kappa], \tag{3.1}$$

be the indicator of whether instance i is "in the top-κ" when ranked according to the predicted values, h. We define two notions of ambiguity in this setting over a dataset sample $S \subset \mathcal{D}$.

Definition 7 (Top-κ ambiguity (all))**.** *The* (ε, κ)-ambiguity (all) *over a sample, S, is the proportion of examples for which the top-*κ *decision changes over the* ε*-Rashomon set:*

$$A_{\varepsilon,\kappa}(h; S) := \frac{1}{|S|} \sum_{i \in S} \max_{h \in \mathcal{H}_\varepsilon(h_0)} 1[Top_{(i,h,\kappa)} \neq Top_{(i,h_0,\kappa)}]. \tag{3.2}$$

Definition 8 (Top-κ Ambiguity (top))**.** *The* (ε, κ)-ambiguity (top) *over a sample, S, is the proportion of top-*κ *examples according to* h_0 *that fall outside the top-*κ *for some models in the* ε*-Rashomon set:*

$$A_{\varepsilon,\kappa}(h; S) := \frac{1}{\kappa} \sum_{i \in S} \max_{h \in \mathcal{H}_\varepsilon(h_0)} Top_{(i,h_0,\kappa)} \left(1 - Top_{(i,h,\kappa)}\right). \tag{3.3}$$

In addition to ambiguity over a sample, we can think about predictive consistency at the individual level. For an individual, we ask whether there is a model in the ε-Rashomon set that can flip the top-κ selection decision. If there is a near-optimal model that flips the top-κ decision, then we can say the point is *flippable*.

Definition 9 (Flippable point)**.** *An instance i is flippable in* $\mathcal{H}_\varepsilon(h_0)$ *if either*

$$Top_{(i,h_0,\kappa)} = 1 \textit{ and } \max_{h \in \mathcal{H}_\varepsilon(h_0)} \tau_i(h) > \kappa \quad ; \textit{or}$$

$$Top_{(i,h_0,\kappa)} = 0 \textit{ and } \min_{h \in \mathcal{H}_\varepsilon(h_0)} \tau_i(h) \leq \kappa.$$

Note that the top-κ ambiguity (all) is simply the fraction of instances that are flippable. Top-κ ambiguity (top) is the fraction of instance in the top-κ of the baseline model h_0 that are flippable out of the top-κ by some $h \in \mathcal{H}_\varepsilon(h_0)$.

3.3.3 On Discrepancy

Recall, Marx et al. [112] introduced a second measure of predictive multiplicity called discrepancy. In this chapter, we focus only on ambiguity because we investigate the overall flexibility in individual predictions over the set of good models. Whereas, discrepancy provides insight into the worst-case scenario by characterizing the maximum proportion of individual flips that would change if the baseline model were replaced with some $h \in \mathcal{H}_\varepsilon(h_0)$. For completeness, we also provide analogous definitions of discrepancy in the resource constrained setting here. But in our methodological details and empirical investigation, we focus only on ambiguity.

Definition 10 (Top-κ Discrepancy (all))**.** *The* (ε, κ)-discrepancy (all) *over a sample, S, is the maximum proportion of examples for which the top-κ decision changes for a model in the ε-Rashomon set:*

$$D_{\varepsilon,\kappa}(h; S) := \frac{1}{|S|} \max_{h \in \mathcal{H}_\varepsilon(h_0)} \sum_{i \in S} 1[Top_{(i,h,\kappa)} \neq Top_{(i,h_0,\kappa)}]. \tag{3.4}$$

Definition 11 (Top-κ Discrepancy (top))**.** *The* (ε, κ)-ambiguity (top) *over a sample, S, is the maximum proportion of top-κ examples according to* h_0 *that fall outside the top-κ for a model in the ε-Rashomon set:*

$$D_{\varepsilon,\kappa}(h; S) := \frac{1}{\kappa} \max_{h \in \mathcal{H}_\varepsilon(h_0)} \sum_{i \in S} Top_{(i,h_0,\kappa)} \left(1 - Top_{(i,h,\kappa)}\right). \tag{3.5}$$

3.3.4 Computing top-κ ambiguity for linear models

In this section we describe the procedure for computing the two notions of top-κ ambiguity for linear models, $\mathcal{H} = \{h(x) = x^T w : w \in \mathbb{R}^{d+1}\}$, and squared error loss, $L(h; \mathcal{D}) = L(w; \mathcal{D}) = RSS(w; \mathcal{D}) = \sum_{i=1}^{n}(y_i - x_i^T w)^2$. We use h and w notation interchangeably in the context of linear models. Unless stated otherwise, we will assume throughout this section that the design matrix X has been transformed to be orthonormal. The problem is invariant to this operation, but working with an orthonormal X helps simplify expressions and reduce notational burden.

We begin by training the *baseline model* h_0 that produces a ranking for each individual in our sample. Our goal is to determine the most meaningful change to this baseline rank for each point over the ε-Rashomon set of competing models. Therefore, for instances with a baseline rank in the top-κ, $Top_{(i,h_0,\kappa)} = 1$, we calculate the *maximum* attainable rank for this individual, $\max_{h \in \mathcal{H}_\varepsilon(h_0)} \tau_i(h)$. For instances with a baseline rank outside the top-κ, $Top_{(i,h_0,\kappa)} = 0$, we calculate the *minimum* attainable rank for this individual, $\min_{h \in \mathcal{H}_\varepsilon(h_0)} \tau_i(h)$. Based on these minimum and maximum ranks, we can compute the proportion of examples whose baseline rank flips over the ε-Rashomon set of competing models.

We employ integer programming for this computation. Prior work involves constructing a pool of candidate models that change individual predictions [112,176]. From that pool of models, those with near-optimal performance are selected to compute ambiguity. These methods are indirect in that the MIPs do not directly constrain these candidate models to be within the ε-Rashomon set. In our setting, we develop a MIP that does include this constraint. The following proposition allows us to neatly characterize the ε-Rashomon set, $\mathcal{H}_\varepsilon(h_0)$, for linear models.

Proposition 1. *Assume the design matrix $X_{n\times(d+1)}$ has been orthonormalized, and w_0 =*

$\text{argmin}_{w \in R^{d+1}} \|y - Xw\|_2^2$ *is the least squares solution. Then*

$$\begin{aligned}\mathcal{H}_\varepsilon(w_0) &= \{w \in \mathbb{R}^{d+1} : RSS(w) \leq RSS(w_0) + \varepsilon\} \\ &= \{w \in \mathbb{R}^{d+1} : \|w - w_0\| \leq \varepsilon\}.\end{aligned}$$

We provide a simple proof, which follows as a corollary of Theorem 10 in Semenova et al. [147].

Proof. Unpenalized linear regression is a special case of ridge regression

$$\min_w L(w; \lambda) = \min_w (y - Xw)^T (y - Xw) + \lambda \|w\|_2^2,$$

with $\lambda = 0$. Part 1 of Theorem 10 of Semenova et al. [147] shows that the ε-Rashomon set for ridge regression is,

$$\mathcal{H}_\varepsilon(w_0; X, \lambda) = \{w : (w - w_0)^T \left(X^T X + \lambda I_{d+1}\right) (w - w_0) \leq \varepsilon\}.$$

For orthonormal designs, $X^T X = I_{d+1}$. This, combined with taking $\lambda = 0$ to recover the unpenalized linear regression setting gives the stated result. □

With this result in hand, we formulate a MIP to calculate the minimum and maximum rank assigned to each point over ε-Rashomon set of competing models, which we call $\text{FlipTopKMIP}(h_0, x_i; \kappa, \varepsilon)$:

$$\begin{aligned}\text{min or} \max_{I \in \mathbb{R}^{2 \times (n-1)}, w \in \mathbb{R}^{d+1}} \quad & \sum_{i' \neq i} I_{(i' > i)} \\ \text{s.t.} \quad &\end{aligned}$$

$$I_{(i'>i)} + I_{(i>i')} = 1 \qquad \forall i' \in S \setminus i \tag{3.6a}$$

$$(x_{i'} - x_i)^T w \leq M * I_{(i'>i)} \; \forall i' \in S \setminus i \tag{3.6b}$$

$$(x_i - x_{i'})^T w \leq M * I_{(i'>i)} \; \forall i' \in S \setminus i \tag{3.6c}$$

$$\|w - w_0\|_2^2 \leq \varepsilon \tag{3.6d}$$

$$w_j \in \mathbb{R} \qquad j = 1, ..., d+1 \tag{3.6e}$$

$$I_{(i'>i)}, I_{(i>i')} \in \{0,1\} \qquad \forall i' \in S \setminus i \tag{3.6f}$$

where we set,

$$M = \left(\sqrt{\|w_0\|_2^2 + \varepsilon}\right) \max_{i,j} \|x_j - x_i\|_2.$$

Now, we provide details on the M bound in constraints (3.6b) and (3.6c). To ensure that $I_{(i'>i)}$ whenever $(x_{i'} - x_i)^T w = \hat{y}(x_{i'}) - \hat{y}(x_i) > 0$ we set M so that

$$M \geq \hat{y}(x_{i'}) - \hat{y}(x_i) \quad \forall i', i, \text{ and } \quad \forall w \in \mathcal{H}_\varepsilon(w_0)$$

Proposition 2.

$$\hat{y}(x_{i'}) - \hat{y}(x_i) \leq \left(\sqrt{\|w_0\|_2^2 + \varepsilon}\right) \max_{i,j} \|x_j - x_i\|_2 \quad \forall i', i, \textit{ and } \quad \forall w \in \mathcal{H}_\varepsilon(w_0)$$

Proof.

$$\max_{w \in \mathcal{H}_\varepsilon(w_0)} \hat{y}_{i'} - \hat{y}_i = \max_{w \in \mathcal{H}_\varepsilon(w_0)} (x_{i'} - x_i)^T w$$

By Cauchy-Schwartz,

$$(x_{i'} - x_i)^T w \leq \|x_{i'} - x_i\|_2 \|w\|_2 \leq \|x_{i'} - x_i\|_2 \max_{w \in \mathcal{H}_\varepsilon(w_0)} \|w\|_2.$$

Noting that

$$\|w\|_2 = \sqrt{\|w_0 + (w - w_0)\|_2^2} \leq \sqrt{\|w_0\|_2^2 + \|(w - w_0)\|_2^2} \leq \sqrt{\|w_0\|_2^2 + \varepsilon} \quad \forall w \in R_\varepsilon(w_0),$$

we therefore get that,

$$M_i = \max_{i'} \max_{w \in R_\varepsilon} \hat{y}_{i'} - \hat{y}_i \leq \left(\sqrt{\|w_0\|_2^2 + \varepsilon} \right) \max_{i'} \|x_{i'} - x_i\|_2.$$

Taking the maximum over all i' gives the desired result,

$$M = \max_{i,i'} \max_{w \in \mathcal{H}_\varepsilon(w_0)} \hat{y}_{i'} - \hat{Y}_i \leq \left(\sqrt{\|w_0\|_2^2 + \varepsilon} \right) \max_{i,i'} \|x_{i'} - x_i\|_2.$$

□

Note that the proof shows that one can set M_i differently for each point i we are aiming to flip in the given run of the MIP.

The objective minimizes (or maximizes) the rank assigned to an individual i. The binary variables $I_{(i'>i)}$ serve as indicators that one point ranks higher than another: $\hat{y}_{i'} = x_{i'}^T w \geq x_i^T w = \hat{y}_i$. So the objective $\sum_{i' \neq i} I_{(i'>i)} = \tau_i(w) - 1$ is simply the rank (minus 1) of point i in model w. Constraint (3.6a) says that between two points, one point has to rank higher or lower making sure there are no ties. We connect the indicators to the rank definition through constraints (3.6b) and (3.6c) where the rank relationship is established using "Big-M" variable M. Constraint (3.6d) enforces that w is in the ε-Rashomon set, as per Proposition (1).

FlipTopKMIP$(h_0, x_i; \kappa, \varepsilon)$ outputs the minimum or maximum rank assigned to each individual in our sample over the ε-Rashomon set. We use this output to determine which points are flippable based on definition 3.4. Then, we simply calculate the proportion of flippable instances to arrive at top-κ ambiguity.

3.3.5 Improving efficiency by identifying provably (un)flippable points

Whereas prior related work on predictive multiplicity in binary [112] and probabilistic [176] classification has involved solving a MIP for every point in $\mathcal{D}$, we show this is not necessary in our setting. Specifically, we show that (i) one can efficiently determine that many points are provably *not flippable* over the ε-Rashomon set; and (ii) one can identify a subset of *flippable* points by solving a proxy optimization problem with a closed-form solution that produces a $w \in \mathcal{H}_\varepsilon(w_0)$ that may flip some points into the top-κ. This means that in practice we only need to solve the computationally expensive FlipTopKMIP for a very small subset of points whose flippability remains undetermined following the two efficient filtering steps. Our approach is grounded in the following three results, whose proofs are below.

Proposition 3 (Prediction gap bound over the ε-Rashomon set)**.** *Define* $\Delta_{i,i'}(w) := \hat{y}_i - \hat{y}_{i'} = x_i^T w - x_{i'}^T w$ *to be the prediction gap between instances* i' *and* i *under model* w*. For all* i, i' *and* $w \in \mathcal{H}_\varepsilon(w_0)$*,*

$$\Delta_{i,i'}(w) \leq \Delta_{i,i'}(w_0) + \sqrt{\varepsilon}\|x_i - x_{i'}\|_2 =: B(i, i'; \varepsilon)$$

Proof of Proposition 3.

$$\begin{aligned}
\Delta_{i,i'}(\boldsymbol{w}) &= x_i^T \boldsymbol{w} - x_{i'}^T \boldsymbol{w} = (x_i - x_i)^\top \boldsymbol{w} \\
&= (x_i - x_i)^\top \boldsymbol{w} + (x_i - x_{i'})^\top \hat{\boldsymbol{w}} - (x_i - x_i)^\top \hat{\boldsymbol{w}} \\
&= (x_i - x_{i'})^\top (\boldsymbol{w} - \hat{\boldsymbol{w}}) + (x_i - x_{i'})^\top \hat{\boldsymbol{w}}
\end{aligned}$$

By Cauchy-Schwartz,

$$\begin{aligned}\left|(x_i - x_{i'})^T(\boldsymbol{w} - \hat{\boldsymbol{w}})\right| &\leq \|x_i - x_{i'}\|_2 \|\boldsymbol{w} - \hat{\boldsymbol{w}}\|_2 \\ &\leq \sqrt{\varepsilon}\|x_i - x_{i'}\|_2,\end{aligned}$$

where in the second step we use the fact that $\boldsymbol{w} \in \mathcal{H}_\varepsilon(w_0)$.

Thus $\forall \boldsymbol{w} \in \mathcal{H}_\varepsilon(w_0)$,

$$\Delta_{i,i'}(\boldsymbol{w}) \leq \Delta_{i,i'}(\hat{\boldsymbol{w}}) + \sqrt{\varepsilon}\|x_i - x_{i'}\|_2 = B(i, i'; \varepsilon).$$

So if $B(i, i'; \varepsilon) < 0$, we have $\Delta_{i,i'}(\boldsymbol{w}) < 0 \; \forall w \in \mathcal{H}_\varepsilon(w_0)$.

□

Corollary 1 (Provably unflippable points)**.** *Suppose i is not in the top-κ for model w_0; i.e., $Top_{(i,w_0,\kappa)} = 0$. If $\#\{i' : B(i, i'; \varepsilon) < 0\} \geq \kappa$, then $Top_{(i,w,\kappa)} = 0 \;\; \forall w \in \mathcal{H}_\varepsilon(w_0)$.*

Proof of Corollary 1. $\{i' : B(i, i'; \varepsilon) < 0)\} \geq \kappa$ means that there are at least κ points for which $\Delta_{i,i'}(\boldsymbol{w}) < 0 \; \forall \boldsymbol{w} \in \mathcal{H}_\varepsilon(w_0)$, so i cannot be in the top-κ set for any model in the ε-Rashomon set. □

Proof of Proposition 4. Let

$$w^* = w_0 + \sqrt{\varepsilon}\frac{x_i}{\|x_i\|_2}.$$

We will show that $\forall w \in \mathcal{H}_\varepsilon(w_0), \hat{y}_i(w) \leq \hat{y}_i(w^*)$. By construction,

$$\hat{y}_i(w^*) = x^T w_0 + \sqrt{\varepsilon}\|x_i\|_2.$$

By Cauchy-Schwartz, for any $w \in \mathcal{H}_\varepsilon(w_0)$

$$\begin{aligned}
\hat{y}_i(w) &= x_i^T w_0 + x_i^T(w - w_0) \\
&\leq x_i^T w_0 + \|x_i\|_2 \|w - w_0\|_2 \\
&\leq x_i^T w_0 + \sqrt{\varepsilon}\|x_i\|_2 \\
&= \hat{y}_i(w^*)
\end{aligned}$$

□

Conceptually, Proposition 3 establishes a bound on the gap between the predicted values of any two points over the whole ε-Rashomon set in terms of the prediction gap under the baseline model, w_0. Corollary 1 then says that if there are at least κ points, $i' \neq i$, whose predicted value is guaranteed to exceed that of point i for every model $w \in \mathcal{H}_\varepsilon(w_0)$, then i is unflippable.

Proposition 4 (Prediction maximizing model)**.** *The predicted value of point i, $\hat{y}_i = x_i^T w$, over the ε-Rashomon set is maximized at,*

$$w^* = \underset{w \in \mathcal{H}_\varepsilon(w_0)}{\operatorname{argmax}} \hat{y}_i(w) = w_0 + \sqrt{\varepsilon}\frac{x_i}{\|x_i\|_2}. \tag{3.7}$$

For points that are not ruled out by Corollary 1, Proposition 4 provides a candidate model within the Rashomon set that may flip a point into the top-κ. Note that this result does not preclude the possibility that $Top_{(i,w^*,\kappa)} = 0$ while also $Top_{(i,w',\kappa)} = 1$ for some other $w' \in \mathcal{H}_\varepsilon(w_0)$. Taken together, these results often significantly reduce the number of points for which one needs to run the MIP in order to determine their flippability.

3.4 Multi-target Multiplicity and Fairness

In the previous section, we introduced the top-κ ambiguity measure for characterizing predictive multiplicity for a single target, y, over the ε-Rashomon set. As discussed at the outset, an important potential source of multiplicity is in the specification of the target outcome itself. In this section, we introduce a measure of multi-target multiplicity for the setting where the multiple targets will ultimately be combined in some way to produce a single score that will be used to prioritize allocation. We also discuss group fairness by examining how the selection rate for a given group varies depending on the specific choice of combining rule.

3.4.1 Multi-target ambiguity and index models

Given candidate targets, $\tilde{y}^{(1)}, \ldots, \tilde{y}^{(K)}$, and features X, we consider a family of "combining procedures," c_α, parameterized by α that map from training data $(X, \tilde{y}^{(1)}, \ldots, \tilde{y}^{(K)})$ to the space of prediction models $\mathcal{H}_\alpha = \{h_\alpha : \mathcal{X} \mapsto \mathbb{R}\}$. Under a resource constraint of κ, resources will then be allocated to the units with the κ highest values of $h_\alpha(x_i)$. We are interested in characterizing how the top-κ set varies across the parameter α governing the combining procedure, c_α. More formally, we define *multi-target ambiguity* as follows.

Definition 12 (Multi-target ambiguity)**.** *The* (α, κ)-multi-target–ambiguity *of a combining procedure* c_α *over a sample* S *is the proportion of examples whose top-*κ *decision varies depending on the choice of* α*,*

$$A_{\alpha,\kappa}(S) := \frac{1}{|S|} \sum_{i \in S} \max_{h_\alpha, h_{\alpha'} \in \mathcal{H}_\alpha} 1\big[Top_{(i,h_\alpha,\kappa)} \neq Top_{(i,h_{\alpha'},\kappa)}\big]. \tag{3.8}$$

Whereas in the single target case we were interested in ambiguity over the ε-Rashomon set, here we focus on ambiguity over the *combining procedure*. Conceptually, a point is "ambiguous" if whether

it is in the top-κ depends on the particular choice of α in the combining procedure. The family of models generated by the combining parameters α is the multi-target family of "good models."

To make the discussion more concrete, we introduce two combining procedures inspired by existing practice, the *index model* approach and the *index variable* approach.

Definition 13 (Index model)**.** *The index model is defined as*

$$\hat{y}_{IM}(x;\alpha) = c_\alpha^{IM}(X, \tilde{y}^{(1)}, \ldots, \tilde{y}^{(K)})(x) = \sum_{k=1}^{K} \alpha_k \hat{y}_k(x), \tag{3.9}$$

where α is a weight vector in the K-simplex, $\alpha \in \mathbb{S}^K := \{\alpha \in \mathbb{R}^K : \sum_{k=1}^K \alpha = 1, \alpha_k \geq 0 \; \forall k\}$, and $\hat{y}_k(x)$ is a prediction model for target $\tilde{y}^{(k)}$.

Note that for this definition to make sense, we assume that the individual predictors $\hat{y}_k$ are first standardized to an appropriate common scale, such as by rescaling $\hat{y} \leftarrow \frac{\hat{y} - mean(\hat{y})}{sd(\hat{y})}$ or converting to percentiles prior to combining. The choice of standardization function does influence results. Choosing a single target outcome k_0 is a special case of an index model with $\alpha_{k_0} = 1$ and $\alpha_k = 0$ for $k \neq k_0$. An advantage of the index model approach is that it places no restrictions on the training procedure used to construct $\hat{y}^{(k)}$. Where appropriate, multi-task learning approaches can be used to jointly learn models across the targets.

This approach is motivated by existing practice in domains such as criminal justice and human services, where multiple so-called scales (i.e., $\hat{y}_k$'s) are constructed to predict different outcomes, and are then aggregated into prioritization schemes or decision recommendations. For instance, the Allegheny Housing Assessment (AHA) tool used to prioritize housing services for persons experiencing homelessness sums the predictions of three $\tilde{y}^{(k)}$ assessed within 12 months of the assessment date: (i) the likelihoood of inpatient mental health services; (ii) the likelihood of jail booking; and (iii) the likelihood of 4 or more ER visits[88].*

*These tools are presented as examples of models that have been constructed for real world applications.

An alternative to index models is an index variable approach, where instead of first forming predictions and then aggregating the different scales, a composite target outcome is formed and then that target is predicted.

Definition 14 (Index variable)**.** *Given candidate targets $\tilde{y}^{(1)}, \ldots, \tilde{y}^{(K)}$, features X, and weights $\alpha \in \mathbb{S}^K$, an index variable model, $\hat{y}_{IV}(x; \alpha)$ is defined by the minimizer,*

$$\hat{y}_{IV}(x; \alpha) = \min_{h \in \mathcal{H}} L(h; \tilde{y}^{(\alpha)}), \quad \textit{where} \quad \tilde{y}^{(\alpha)} = \sum_{k=1}^{K} \alpha_k \tilde{y}^{(k)}. \tag{3.10}$$

Conceptually, the index variable approach can be thought of as first forming a composite outcome that is believed to more comprehensively describe some latent quantity, and then finding the optimal predictor for that outcome. Both for the index model and index variable formulation, the parameter α captures potential underspecification in the choice of target. In the case of linear models, the index model and index variable approach coincide.

Proposition 5 (Equivalence of index model and index variable approaches for linear models.)**.** *If we restrict consideration to linear models whose solution takes the form $\hat{y} = M_X y$ for some $n \times n$ matrix M_X that depends on X but not on y, then the index model and index variable approach are equivalent.*

Proof of Proposition 5. Starting with the index variable definition, we get that

$$\hat{y}_{IV}^{(\alpha)} = M_X \tilde{y}^{(\alpha)} = M_X \left(\sum_{k=1}^{K} \alpha_k \tilde{y}^{(k)} \right) = \sum_{k=1}^{K} \alpha_k M_X \tilde{y}^{(k)} = \sum_{k=1}^{K} \alpha_k \hat{y}^{(k)} = \hat{y}_{IM}^{(\alpha)}$$

□

Note that linear regression is a special case of a linear model, with $M_X = X(X^T X)^{-1} X^T$. Other models such as regression splines fall into this class as well.

We are not endorsing the use of these other tools.

In the remainder of this work we focus on the index model approach, as it can be analysed in a computationally tractable way for general predictors $\hat{y}_k$. Due to the equivalence result, our methods are directly applicable to the index variable approach in the case of linear models.

3.4.2 Computing multi-target top-κ ambiguity for index models

In this section, we introduce a MIP for computing multi-target ambiguity as defined in Eq. (3.8) for the family of index models. The MIP calculates the minimum and maximum rank attainable for each individual point over the combining parameters, α. The multi-target ambiguity is then given as the proportion of points for which the minimum rank is $\leq \kappa$ while the maximum rank $\geq \kappa$.

For combining procedures parameterized by α, the min and max rank of each individual $i \in S$ can be obtained by solving the optimization problem, which we call $\mathrm{FlipTopKMultiMIP}(x_i; \kappa)$:

$$\min_{I \in \{0,1\}^{n-1},\, \alpha \in \mathbb{R}^K} \sum_{i' \neq i} I_{(i'>i)} - 0.5 \sum_{k=1}^{K} \alpha_K \text{ or}$$
$$\max_{I \in \{0,1\}^{n-1},\, \alpha \in \mathbb{R}^K} \sum_{i' \neq i} I_{(i'>i)} + 0.5 \sum_{k=1}^{K} \alpha_K$$

s.t.

$$I_{(i'>i)} + I_{(i>i')} = 1 \qquad i' = 1, ..., n \setminus i \tag{3.11a}$$
$$\hat{y}_{IM}(x_{i'}; \alpha) - \hat{y}_{IM}(x_i; \alpha) \leq M * I_{(i'>i)} \qquad i' = 1, ..., n \setminus i \tag{3.11b}$$
$$\hat{y}_{IM}(x_i; \alpha) - \hat{y}_{IM}(x_{i'}; \alpha) \leq M * I_{(i>i')} \qquad i' = 1, ..., n \setminus i \tag{3.11c}$$
$$0 \leq \alpha_k \leq 1 \qquad k = 1, ..., K \tag{3.11d}$$
$$0.1 \leq \sum_{k=1}^{K} \alpha_k \leq 1 \tag{3.11e}$$
$$\alpha_k \in \mathbb{R} \qquad K = 0, ..., d \tag{3.11f}$$
$$I_{(i'>i)}, I_{(i>i')} \in \{0, 1\} \qquad i' = 1, ..., n \setminus i \tag{3.11g}$$

where $\hat{y}_{IM}(x_i; \alpha)$ is shorthand for $\sum_{k=1}^{K} \alpha_k \hat{y}_k(x_i)$, and

$$M = \max_{i',k} \hat{y}^{(k)}(x_{i'}) - \min_{i,k} \hat{y}^{(k)}(x_i).$$

FlipTopKMultiMIP$(x_i; \kappa)$ fits the parameters α that minimize (or maximize) the rank assigned to individual i. The objective minimizes (or maximizes) the sum of individuals ranked higher than individual i with an additional term in the objective that forces $\sum \alpha_k = 1$. Again, the binary variables $I_{(i'>i)}$ serve as indicators that one point ranks higher than another. And constraint (3.11a) says one point has to rank higher or lower than another (i.e. there are no ties). We connect the indicators to the ordering relations $\hat{y}_{IM}(x_{i'}; \alpha) \geq \hat{y}_{IM}(x_i; \alpha)$ and $\hat{y}_{IM}(x_{i'}; \alpha) < \hat{y}_{IM}(x_i; \alpha)$ through constraints (3.11b) and (3.11c), introducing the "Big-M" variable, M. Constraints (3.11d) and (3.11e) ensure are a soft version of the constraint that α is in the simplex, $\mathbb{S}^K$.

This formulation is similar to FlipTopKMIP from the single-target case, but the objective has an additional term, and the optimization here is over the combining weights α rather than the parameters of the individual predictors $\hat{y}_k$. As in the single target setting, we calculate ambiguity by identifying flippable points using a MIP. In this setting, there is no baseline model, so the term "flippable" now refers to points where there exist two choices of combining parameters, $\alpha \neq \alpha'$, such that $Top_{(i,\alpha,\kappa)} \neq Top_{(i,\alpha',\kappa)}$. Furthermore, the optimization here is no longer over an ε-Rashomon set—a notion which does not naturally extend to the multiple target setting due to the absence of a baseline model—but rather over the parameters α governing the combining rule.

As in the single-target context, we can once more reduce the number of times we need to run the MIP by identifying points that provably cannot appear in the top-κ set for any choice of α, and characterize the prediction-maximizing choice of α for each point. The results and accompanying proofs are in § 3.4.3.

3.4.3 Identifying certifiably (un)flippable points in the multi-target setting without solving a MIP

Here, we provide more technical details for identifying (un)flippable points.

Proposition 6 (Prediction gap bound for index models.)**.** *Let* $\Delta_{i,i'}(\alpha) := \hat{y}_{IM}(x_i;\alpha) - \hat{y}_{IM}(x_{i'};\alpha)$ *to be the prediction gap between instances* i' *and* i *under combining parameters* α*. For all* i, i' *and* $\alpha, \alpha' \in \mathbb{S}^K$*,*

$$\Delta_{i,i'}(\alpha) \leq \Delta_{i,i'}(\alpha') + \sum_{k=1}^{K} |\hat{y}_k(x_i) - \hat{y}_k(x_{i'})| =: B_{IM}(i, i'; \alpha).$$

Proof. For any two instances $x_i, x_{i'} \in \mathcal{X}$ and combining parameter vectors $\alpha, \alpha' \in \mathbb{S}^K$,

$$\begin{aligned}
\Delta_{i,i'}(\alpha) &= \Delta_{i,i'}(\alpha') + \sum_{k=1}^{K} (\alpha_k - \alpha'_k)\left(\hat{y}_k(x_i) - \hat{y}_k(x_{i'})\right) \\
&\leq \Delta_{i,i'}(\alpha') + \sum_{k=1}^{K} |\hat{y}_k(x_i) - \hat{y}_k(x_{i'})| \\
&= B_{IM}(i, i'; \alpha)
\end{aligned}$$

□

Corollary 2 (Points that cannot appear in top-κ set for any index model)**.** *Suppose* i *is not in the top-*κ *for an index model with parameter* $\tilde{\alpha}$*; i.e.,* $Top_{(i,\tilde{\alpha},\kappa)} = 0$*. If* $\#\{i' : B(i, i'; \tilde{\alpha}) < 0\} \geq \kappa$*, then* $Top_{(i,w,\kappa)} = 0 \ \forall \alpha \in \mathbb{S}^K$*.*

Proof. $\{i' : B_{IM}(i, i'; \tilde{\alpha}) < 0)\} \geq \kappa$ means that there are at least κ points, i', for which $\Delta_{i,i'}(\alpha) < 0 \, \forall \alpha \in \mathbb{S}^K$, so i cannot be in the top-κ set for any index model. □

Proposition 6 establishes a bound on the gap between the predicted values of any two points *for all* $\alpha \in \mathbb{S}^K$ in terms of the prediction gap under *any one* choice of combining parameters α.

Corollary 2 then allows us to determine when i cannot be in the top-κ of *any* index model $\alpha \in \mathbb{S}^K$ based on the prediction gap for a *given* $\tilde{\alpha}$.

Proposition 7 (Prediction maximizing index model)**.** *The predicted value of point i is maximized at $\alpha^* \in \mathbb{S}^K$ where $\alpha^*_{k^*} = 1$ for $k^* = \text{argmax}_k \hat{y}_k(x_i)$ and $\alpha^*_k = 0$ for $k \neq k^*$.*

Proof.

$$\max_{\alpha \in \mathbb{S}^K} \hat{y}^{(\alpha)}_{IM}(x_i) = \max_{\alpha \in \mathbb{S}^K} \sum_{k=1}^{K} \alpha_k \hat{y}_k \leq \max_k \hat{y}_k \sum_{k=1}^{K} \alpha_k = \max_k \hat{y}_k,$$

which is achieved at the stated value of the combining parameter vector, α^*. □

Proposition 7 provides a candidate α for which a given point may be in the index model's top-κ. Note that this result does not preclude the possibility that $Top_{(i,\alpha^*,\kappa)} = 0$ while also $Top_{(i,\alpha',\kappa)} = 1$ for some other $\alpha' \in \mathbb{S}^K$. This result suggests the simple strategy of first identifying points whose top-κ decision varies between the single-target prediction models $\hat{y}_k$.

3.4.4 Group-level selection rates in top-κ selection with multiple targets

In this section we demonstrate how our framework can be applied to examine group fairness concerns. Specifically, we consider how the selection rate—i.e., the proportion of instances from a given group, $A = a$, in the top-κ—varies with the combining weights α.[†]

We can compute the combining parameters that maximize the number of individuals in a given group who are selected to be in the top-κ. That is, we consider,

$$\min_\alpha \textit{or} \max_\alpha \sum_{i=1}^{n} 1[A = a] Top_{(i,\alpha,\kappa)} = \min_\alpha \textit{or} \max_\alpha \sum_{i \in G_a}^{n} Top_{(i,\alpha,\kappa)} \tag{3.12}$$

[†]While, in principle, one can also examine measures such as the False Positive Rate and True Positive Rate by analysing the subsample of instances for which $\tilde{y}^{(k)} = 0$ (or 1, for TPR), it is not entirely clear how such quantities should be interpreted. How should one weigh a high FPR for a given target against a low FPR for a different one in a setting where the "correct" choice of target is itself in doubt?

where $G_a = \{i : A_i = a\}$ denotes all the instances that are in protected group $A = a$. While the goal of our work is to characterize the variation in selection rates afforded by different choices of combining parameters, α, the methods can also be used to select a particular model that maximizes (or minimizes) those rates.

We compute the quantities in Eq. (3.12) through another MIP. For this purpose we introduce variables $T_i \in \{0,1\}$ that play the role of the $Top_{(i,\alpha,\kappa)}$ indicator. We refer to this MIP as

GroupSelectRateTopKMultiMIP$(a;\kappa)$:

$$\min_{I\in\{0,1\}^{2n\times|G_a|},\ T\in\{0,1\}^{|G_a|},\ \alpha\in\mathbb{S}^K} \quad \sum_{i\in G_a} T_i - 0.5\sum_{k=1}^{K}\alpha_k \text{ or}$$

$$\max_{I\in\{0,1\}^{2n\times|G_a|},\ T\in\{0,1\}^{|G_a|},\ \alpha\in\mathbb{S}^K} \quad \sum_{i\in G_a} T_i + 0.5\sum_{k=1}^{K}\alpha_k$$

s.t.

$$I_{(i'>i)} + I_{(i>i')} = 1 \qquad \forall i \in G_a, \forall i' \in S \setminus i \tag{3.13a}$$

$$\hat{y}_{IM}(x_{i'};\alpha) - \hat{y}_{IM}(x_i;\alpha) \le M_I * I_{(i'>i)} \quad \forall i \in G_a, \forall i' \in S \setminus i \tag{3.13b}$$

$$\hat{y}_{IM}(x_i;\alpha) - \hat{y}_{IM}(x_{i'};\alpha) \le M_I * I_{(i>i')} \quad \forall i \in G_a, \forall i' \in S \setminus i \tag{3.13c}$$

$$\kappa - \sum_{i'\neq i} I_{(i'>i)} \le \kappa * T_i \qquad i \in G_a \tag{3.13d}$$

$$\left(1 + \sum_{i'\neq i} I_{(i'>i)}\right) - \kappa \le (n-\kappa)(1-T_i) i \in G_a \tag{3.13e}$$

$$0 \le \alpha_k \le 1 \qquad k = 1,\ldots,K \tag{3.13f}$$

$$0.1 \le \sum_{k=1}^{K}\alpha_k \le 1 \tag{3.13g}$$

$$I_{(i'>i)}, I_{(i>i')} \in \{0,1\} \qquad i \neq i' = 1,\ldots,n \tag{3.13h}$$

$$T_i \in \{0,1\} \qquad i \in G_a \tag{3.13i}$$

The objective minimizes (or maximizes) the number of individuals in group G_a that are selected to be in the top-κ. There is an additional term in the objective that forces $\sum \alpha_k = 1$, which has the effect of enforcing the simplex constraint on α. Recall, the binary variables $I_{(i'>i)}$ serve as indicators that one point ranks higher than another. Thus, constraint (3.13a) means that one point has to rank higher or lower making sure there are no ties. We connect the indicators to the ordering relations $\hat{y}_{IM}(x_{i'};\alpha) \geq \hat{y}_{IM}(x_i;\alpha)$ or $\hat{y}_{IM}(x_{i'};\alpha) < \hat{y}_{IM}(x_i;\alpha)$ through constraints (3.13b) and (3.13c) using the "Big-M" variable M_I. This value is set to be the max possible difference in prediction between two points $M_I = \max_{i,k} \hat{y}^{(k)}(x_i) - \min_{i,k} \hat{y}^{(k)}(x_i)$. To make sure T_i reflects whether individuals are in group G_a and ranked in the top-κ, we have constraints (3.13d) and (3.13e). Constraints (3.13f) and (3.13g) are other pieces of the simplex constraint.

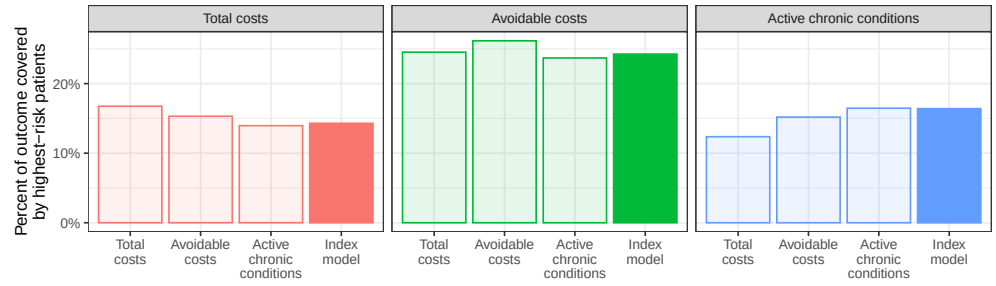
3.5 Stable points

Thus far our focus has been on multiplicity and identifying *flippable* points—those whose decision depends on the particular model chosen among the set $\mathcal{G}$ of good models.[‡] In practice, however, we may be equally interested in *un*flippable points. As prior work has pointed out, the presence of multiplicity raises concerns about arbitrariness: What justification can you offer someone who receives an adverse decision from the chosen model when there may exist another good model that would have given them a favorable decision [22]? Our work can speak to this as well. Concretely, our proposed methods can be used to identify what we call *stable points*: cases whose decisions do not change over the set of good models.

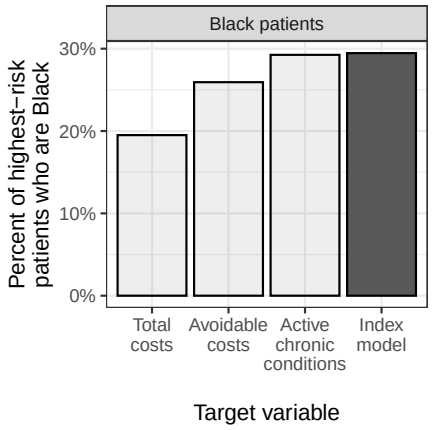
Definition 15. *Let $\mathcal{G}$ be the set of "good models". We say that i_0 is a stable κ-selected point if $Top_{(i_0,h,\kappa)} = 1 \; \forall h \in \mathcal{G}$. Similarly, we say that i_0 is a stable κ-unselected point if $Top_{(i_0,h,\kappa)} = 0 \; \forall h \in \mathcal{G}$.*

[‡] E.g., $\mathcal{G} = \mathcal{H}_\varepsilon(h_0)$ in the single target setting, or the set of index models parameterized by $\alpha \in \mathbb{S}^K$ in the multi-target setting.

Stable points are instances for which the decision is non-arbitrary: their decisions are invariant to the specific choice of model among those considered acceptable, which is strong justification for the given decision. The fraction of stable κ-selected points out of κ is a useful quantifier of the arbitrariness of a predictive allocation task. For instance, if this fraction is very low, this may highlight a need for further principled deliberation on the specific choice of model, or affect the willingness to adopt a predictive model for the given allocation task.



(A)



(B)

Figure 3.1: (A) The concentration of various outcomes under models optimized for different targets. Each panel shows the percent of an outcome captured by the highest-risk patients relative to the entire outcome distribution across all patients. Each bar represents one type of model. The transparent bars depict models trained to predict individual targets, whereas the solid bars depict the index model, which re-weights the individual predictions to maximize fairness. (B) The percent of Black patients among highest-risk patients identified by each model.

3.6 Evaluation

In this section, we apply the techniques developed above to the healthcare dataset analyzed by Obermeyer et al. to better understand the opportunities afforded by multiplicity among multiple target variables. First, we describe the dataset in more detail and then apply our multi-target multiplicity framework to it. We then construct a semi-synthetic version of this dataset to develop intuition for the conditions under which we should (or should not) expect to see gains from index models. Finally, we compare the degree of multiplicity that arises in resource constrained settings under a single target to the same under multiple targets. Throughout this section, we solve all integer programs with Gurobi v.9.5.2 [68]. Our software implementation is at https://github.com/JWatsonDaniels/multitarget-multiplicity.

3.6.1 Dataset

We demonstrate our framework on a dataset released by Obermeyer et al., which is unique in several ways. The original paper examines patient data for all primary care patients at a large academic hospital. However, due to the sensitivity of the data, the authors were unable to release the dataset in its original form. Instead, they created a publicly available semi-synthetic version of the dataset that is designed to closely mirror the original dataset.[§]

The released dataset contains several related but different outcomes for patients in a given year including total healthcare costs, avoidable healthcare costs (emergency visits and hospitalizations), and number of active chronic illnesses. It also contains a rich set of features about each patient, including demographics (age, sex, race) and information about the patient's health and healthcare costs in the previous year. Specifically, there are indicators for individual chronic illnesses that a patient had in the previous year, costs claimed by the patients' insurer in the previous year, biomarkers for

[§] https://gitlab.com/labsysmed/dissecting-bias

medical tests from the previous year, and medications taken in the previous year.

In the original paper, Obermeyer et al. examine a proprietary scoring system used by the hospital to identify high-risk patients. The risk scores are generated by a model designed to predict healthcare costs in the current year based on patient demographics and healthcare information available from the previous year. In particular, patients who are assigned risk scores that fall in the 97th percentile or above (i.e., the top 3% of assigned scores) are automatically identified for inclusion in the hospital's "high-risk care management" program.

The authors examine the assigned risk scores in detail and show that they contain a significant racial bias. Specifically, they find that Black patients at a given risk score have worse health outcomes, on average, than their White counterparts. The authors trace this bias back to the choice of predicting healthcare costs as the target variable. Due to differences in access to healthcare, White patients tend to have higher healthcare costs, on average, than Black patients of similar health. This difference is then reflected in the developed risk score, leading to the observed racial bias. Obermeyer et al. then go on to show that there are different target variable choices that exhibit less of a racial bias—specifically using either avoidable costs or active chronic illnesses as a target instead of total costs.

3.6.2 Optimizing across healthcare outcomes

We present a re-examination of this healthcare dataset to further explore the ways in which flexibility in target variable choice can be used to address fairness concerns. Obermeyer et al. consider using one of each of the three different target variables, which in our framework corresponds to an index model with binary α weights. For example, the cost model can be thought of as $\hat{y}_{IM} = 1 \cdot \hat{y}_{\text{costs}} + 0 \cdot \hat{y}_{\text{avoidable costs}} + 0 \cdot \hat{y}_{\text{active illnesses}}$. However, these are just three extremes among the possible set of index models that can be formed with a continuous α to create a weighted average of the three available target variables.

Our analysis explores whether exercising these extra degrees of freedom can lead to more equi-

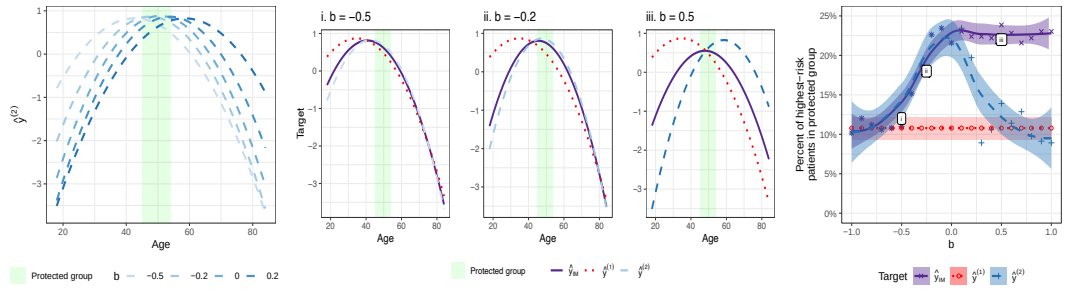


Figure 3.2: (A) A semi-synthetic family of models $\hat{y}_2$, which go from negatively correlated with age $(b < 0)$ to positively correlated with age $(b > 0)$. Patients in the protected group are concentrated in the middle age range, indicated by the green band. (B) A closer look at all three models for three different values of b. In the left panel, none of the models peak in the middle age range. In the middle, the index model $\hat{y}_{IM}$ (solid purple) learns to ignore $\hat{y}_1$ (dotted red) in favor of $\hat{y}_2$ (dashed blue) to capture more of the protected group. On the right, neither of the individual models peak to capture the protected group, but the index model averages them to do so. (C) A more detailed look at the concentration of the protected group found by each model over the range of b values, showing that the index model dominates either individual model over the entire range.

table outcomes. To address this, we replicate and extend the analysis in Table 2 of the original paper, using the released dataset.[¶] Specifically, we train separate models to predict each of the three target variables (healthcare costs, avoidable costs, and active chronic illnesses) and use the fitted models to rank a held-out set of patients.[‖]

We identify the top 3% of highest-risk patients according to each of the models and look at the concentration of outcomes and the racial composition of the identified patients. For instance, when considering total costs, we compute what percent of all costs (across all patients) are covered by just the highest-risk patients. When considering active chronic illnesses, we instead compute the fraction of all illnesses (across all patients) covered by this set. We then extend these results by running the multi-target fairness mixed-integer programming (MIP), GroupSelectRateTopKMultiMIP, to search for an index model that maximizes the fraction of Black patients concentrated among the

[¶] The original table is generated using the proprietary, unreleased data. Replicating the table with the released dataset produces similar, but not identical results for this reason.

[‖] We use the train/holdout set specified by the authors in the released dataset. We train OLS linear regression models for each variable. In order to do so, we remove several co-linear features provided in the released dataset, detailed in § 3.6.5.

highest-risk set. Figure 3.3 outlines more details of this process.

The results are displayed in Fig. 3.1 and show several key observations. First, we see that, as expected, the model trained to predict a given individual outcome has the largest concentration of that outcome in the high-risk patient set. Notice that the transparent bars on the far left panel of Fig. 3.1A show that the model trained to predict total cost is the one that has the highest concentration of total costs in the high-risk patient set. Conversely, on the far right panel we see that modeling active chronic conditions produces the highest concentration of current illnesses in the high-risk set. Second, despite these differences we see comparatively small variation in outcome concentration across different target variable choices, with less than a 5 percentage point difference across models in the first three panels. But, we do see a substantial difference in the racial composition of the high-risk set, as indicated in Fig. 3.1B—a more than 10 percentage point difference.

The index model, in comparison, is shown in the solid bars of Fig. 3.1 for $\alpha = (0.05, 0.0, 0.95)$. By comparing the solid bars to the transparent ones, we see that the index model does a reasonable job of capturing each of the individual targets that it is comprised of, but also produces a high-risk set with a high concentration of Black patients, as per the objective of the multi-target group selection formulation (3.12). In effect, this represents a "best of both worlds" solution: we are able to fit separate models that are useful for predicting the three outcomes that may be of interest on their own (i.e., for budgeting purposes), but we also arrive at a way of ranking patients that results in a more equitable allocation of a scarce resource via the index model.

3.6.3 Exploring the conditions for effective multi-target optimization

In the example above, we found it was possible to learn an index model that combined individual target variables from the healthcare dataset to improve group selection rates. In this section, we use semi-synthetic data to gain a better understanding of the conditions for which we might (or might not) expect to see such gains in other datasets. To do this, we modify the healthcare dataset

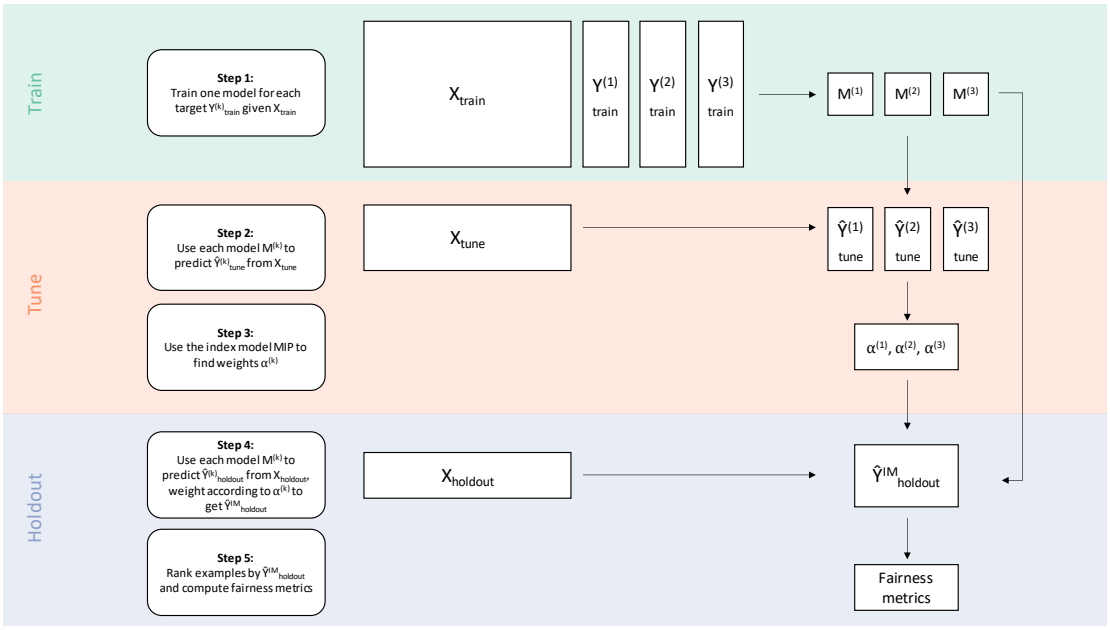
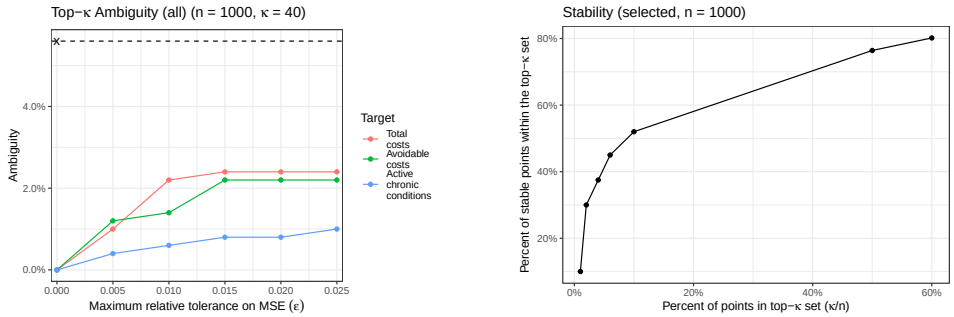


Figure 3.3: Workflow for maximizing fairness with multiple targets. Training data is used to fit separate models for each target variable. In the tune phase, the fitted models are used to forecast each target variable, and the index model MIP is run to find a fairness-maximizing weighted combination of the targets. Finally, in the holdout phase a separate dataset is used to calculate the weight index model predictions, from which fairness metrics are computed.

to systematically control the relationship between a protected group attribute a, a feature x, and the different choices of target $\tilde{y}^{(k)}$. We then vary these relationships and examine how this affects the group selection rate that an index model can achieve.

Specifically, we construct a dataset with one feature (age) and two target variables $\tilde{y}^{(1)}$ and $\tilde{y}^{(2)}$, along with a protected attribute.** We construct the protected attribute to be non-monotonically correlated with age, with a higher concentration of patients in the protected group falling in the middle age range compared to the rest of the population. We then construct one target variable $\tilde{y}^{(1)}$ that is negatively correlated with age and one target variable $\tilde{y}^{(2)}$ whose correlation with age varies from strongly positive to strongly negative, controlled by a parameter b, as shown in Fig. 3.2A. In

**While age is considered a protected attribute under various discrimination laws, for the purposes of our evaluation in a healthcare setting, we treat it as an unprotected attribute.



(A) **(B)**

Figure 3.4: (A) Comparison of multiplicity within vs. between targets. Ambiguity within each individual target is shown by the colored lines at different relative mean squared error tolerances (ε). Ambiguity across the three targets, shown by the black 'x' and dotted line, is much higher than ambiguity for any individual target. (B) Stability of points within the top-κ set as κ is increased. Even with relatively small values of κ, we find a sizeable set of stable points that, no matter how targets are combined, fall in the top-κ set.

this setting, prioritizing middle-aged patients maximizes the fraction of high-risk set that is in the protected group, but fitting a model to $\tilde{y}^{(1)}$ prioritizes young patients, resulting in a lower selection rate. Conversely, when b is large and positive, $\tilde{y}^{(2)}$ is positively correlated with age, and so fitting a model to it will prioritize older patients, also leading to sub-optimal group selection. However, as we show in Figs. 3.2B and 3.2C, an index model can be fit over a wide range of b values such that the group selection rate is maximized. The intuition is that the index model can learn to average out unhelpful correlation structure between the protected attribute and the target variables.

3.6.4 Multi-target versus individual-target multiplicity

Finally, we compare the latitude afforded by across-target multiplicity to that for within-target multiplicity. To do this we return to the original dataset released by Obermeyer et al. and work with a subset of the features and examples for computational efficiency, as described in § 3.6.5.

We evaluate predictive multiplicity by computing the single-target top-κ ambiguity for each choice of target variable Eq (3.3) by running FlipTopKMIP for different error tolerances ε. This allows us to determine the proportion of top-κ points that can be flipped. The results are shown in

Fig. 3.4A, with each color corresponding to a choice of target variable. From this, we see that single-target ambiguity rises quickly with ε and then plateaus. This is a result of the resource constrained predictive allocation setting: at a certain level, ε, the ε-Rashomon set contains the "flipping" model for each of the flippable points, so further increasing ε does not further increase ambiguity. We observe that the total cost variable has the highest ambiguity, slightly above 2%, whereas active chronic conditions plateau just above 1%.

We compute the multi-target top-κ ambiguity (3.8) by running FlipTopKMultiMIP across the three different target variables. This results in a multi-target ambiguity of nearly 5%, as indicated by the black "x" and dashed horizontal line in Fig. 3.4A. From this we see that the across-target multiplicity is substantially higher than the within-target multiplicity—a much higher proportion of points can be flipped into the top-κ set by re-weighting predictions for the different targets than by entertaining slightly sub-optimal model fits for the individual targets.

Finally, in Fig. 3.4B we look at the complement of ambiguity, examining the set of stable points that remain in the top-κ set over all possible index models, as defined in §3.5. Specifically, we plot the percent of stable points in the top-κ set as we increase κ to cover more of the entire dataset. Interestingly we see that the fraction of stable points grows rapidly with κ. For instance, when combining individual targets under an index model to select the top 10% of highest risk patients, more than half of the selected patients are the same *regardless of which index model is used*. This invariance lends confidence to the decision to prioritize such patients.

3.6.5 Dataset details

We remove several co-linear features from the original Obermeyer et al. dataset so that models can be fit with OLS regression (instead of regularized regression). Specifically, we remove features whose

variable name matches the following regular expression:[††]

```
(gagne_sum_tm1|normal_tm1|esr_.*-low_tm1|
crp_(min|mean|max).*_tm1|ghba1c_.*-low_tm1)
```

This eliminates the sum of active illnesses (which are listed as individual binary features in the dataset) as well as one-hot encoded indicators for individual test results that have low/normal/high levels.

For §3.6.3 we use a smaller subset of features, for computational efficiency, taking only features that match the following regular expression:

```
(gagne_sum_tm1|hypertension_elixhauser_tm1|^dem_|cost.*tm1)
```

This takes the count of total illnesses in the previous time period instead of individually coded illnesses along with demographics and cost in the previous time period.

3.7 Concluding Remarks

In this paper, we introduced frameworks for assessing the level of multiplicity present in predictive allocation tasks in both the single-target and multi-target setting. First, we show how to measure multiplicity for a given target variable in settings where decision makers face constraints that limit the total number of people who can receive a scarce resource. Second, we show that when faced with a choice of multiple target variables, practitioners can develop index models that address fairness concerns by re-weighting and combining predictions for each target. Our empirical results show

[††]See data dictionary for names and descriptions of variables here: https://gitlab.com/labsysmed/dissecting-bias/-/blob/master/data/data_dictionary.md

that both of these methods are effective for narrowing racial disparities in selection rates in allocating patients to a high-risk coordinated care management program. Notably, we find that the latitude afforded by re-weighting predictions across target variables is substantially larger than the flexibility provided by leveraging within-target multiplicity. This may represent a "best of both worlds" solution: we are able to fit separate models for predicting outcomes that may be interesting to model in their own right, but we can also combine the predictions from these models to allocate resources more equitably.

Systems do not maintain themselves; even our lack of intervention is an act of maintenance. Every structure in every society is upheld by the active and passive assistance of other human beings.

Sonya Renee Taylor

4

Predictive Churn with the Set of Good Models

4.1 Introduction

One of the foremost challenges faced in the deployment of machine learning (ML) models used in consumer-facing applications is unexpected changes over periodic updates. Model updates are essen-

tial practice for maintaining and improving long-term performance in mass-market applications like recommendation and advertising. In applications like credit scoring and clinical decision support, however, changes in individual predictions may lead to inadvertent effects on customer retention and patient safety.

At the same time, prediction stability – i.e., consistent, reliable, and predictable behavior – is a basic expectation of ML models used to support human decision-making. Hence, a challenge in ML practice is guaranteeing the stability of predictions made by deployed models after they are updated. Examples of model updates that may impact individual-level predictions include updating parameters via additional training steps on new data, adding input features, and quantizing weights[75,33,37,140].

Unexpected or unreliable predictions after an ML model update can illicit safety concerns when models influence human decision-making. For instance, many clinicians use risk models to support a range medical decisions, from diagnosis to prognosis to treatment[122,160,85]. Updates to a medical model, though potentially rendering better *average* performance, may fundamentally impact the treatment selected for individual patients. As another example, lenders also use risk models to support financial decision-making, i.e., predicting the risk that a borrower will fail to make payments or default on a loan[5,9]. Here, instability after a model update can lead to loan denials to applicants who previously would have been approved – even if the new model is more accurate on average.

In both examples, patients or borrowers impacted by inconsistent predictions merit further analysis to avoid arbitrary, harmful, and unfair decisions. Hence, a number of methods aim to ensure that predictions do not change significantly after a model update – only enough to reflect an average gain in predictive accuracy. For instance, recent work in interpretable ML imposes a "maximum deviation" constraint to control how far a supervised learning model deviates from a 'safe' baseline[177]. The idea is that a significant deviation from expected behavior is problematic. Therefore, methods for assessing this type of predictive (in)stability are a means by which to examine safety. This chapter

focuses on exploring the relationship between two facets of (in)stability in applied ML: predictive churn and predictive multiplicity.

Predictive Churn considers the differences in individual predictions between models pre- and post-update. Predictive churn is formulated in terms of two models: a current deployed model, and an updated model resulting from training the current model on additional fresh data [38]. In several applications, a high level of predictive churn is undesirable. For example, in loan approval, predictive churn can lead to inconsistent applicant experiences (a loan previously granted being denied post-model update).

Predictive Multiplicity occurs when models that are "equally good" on average (e.g., achieve comparable test accuracy) assign conflicting predictions to individual samples [112]. Several recent works demonstrate that many ML tasks admit a large *Rashomon Set* [56,47] of competing models that can disagree on a significant fraction of individual predictions [176,76,97,174]. In ML-supported decision-making, the arbitrary selection of a model from the Rashomon Set without regard for predictive multiplicity can lead to unjustified and arbitrary decisions [22,40]. Within the literature on predictive multiplicity, the Rashomon Set can be defined with respect to an optimal (baseline) model or without. In this chapter, the Rashomon set defined without respect to a baseline is a set of models with similar performance derived from varying random seed initialisations. This is distinct from the Rashomon set considered in the two previous chapters.

The concepts are distinct in their motivation and methodological study. Model Multiplicity tends to have motivations in fairness and interpretability. Meanwhile, predictive churn is motivated by practical industry concerns about periodic model updates. They both involve slight perturbations in the training pipeline though one from perturbing training data directly and the other from perturbing the model directly. Intuitively, I imagine instances where a dataset perturbation might be equivalent to a model perturbation. Similarly, there may be instances where they differ in important ways. This chapter explores this relationship more closely. The main contributions include:

1. We examine whether individual predictions that are unstable under model perturbations (multiplicity) are also those that are unstable under dataset perturbations (churn). We compare between examples that are ε-Rashomon *unstable* and *churn unstable*. For a fixed test sample, we find that the ε-Rashomon *unstable* set does often contain most examples within the *churn unstable* set. The proportion of *churn unstable* examples included in the Rashomon *unstable* ranges from 50% to 100% across datasets and model types for a small dataset update. Results also show that when a model exhibits high predictive multiplicity on one dataset relative to others, it also exhibits high predictive churn (across both churn regimes) relative to other datasets. In practice, analyzing predictive multiplicity (via an empirical ε-Rashomon set) can help anticipate the severity of predictive churn over future model updates.

2. We theoretically characterize the expected churn between models within the ε-Rashomon set from different perspectives. Our analysis reveals that the potential for reducing churn by substituting the deployed model with an alternative from the ε-Rashomon set hinges on the training procedure employed to generate said ε-Rashomon set. The results also show that when updating from model A to model B, we can produce both ε-Rashomon sets (with respect to a baseline) and analytically compute an upper bound on the churn between them.

3. We present empirical results showing that analyzing predictive multiplicity is useful for anticipating churn even when a model has been enhanced with uncertainty awareness. We question whether models with inherent uncertainty quantification abilities might (i) exhibit less predictive multiplicity and (ii) produce individual uncertainty estimates that indicate which examples will be ε-Rashomon or churn unstable. Our findings show that in fact there can be more predictive multiplicity for an uncertainty aware (UA) model though the uncertainty estimates do prove helpful in anticipating unstable instances from either perspective.

4.2 Related Work

Model Multiplicity Model multiplicity in machine learning often arises in the context of model selection, where practitioners must arrive at a single model to deploy [30,25], from amongst a set of near-optimal models, known as the "Rashomon" set. There are a number of studies focused on examining the Rashomon set [56,47,147,185,48]. Predictive multiplicity is the prevalence of *conflicting* predictions over the Rashomon set and has been studied in binary classification [112], probabilistic classification [176,76], differentially private training [97] and constrained resource allocation [174]. There is a growing body of research on the implications of differences in models within the Rashomon set [51,168,134,40,19,1,22,105] and on predictive arbitrariness and randomness in a more general setting [36,119,60]. Distinctively, the present chapter applies the Rashomon perspective to uncover insights about predictive churn.

Predictive Churn Predictive churn is a growing area of research. Cormier et al. [38] define churn and present two methods of churn reduction: modifying data in future training, regularizing the updated model towards the older model using example weights. Churn reduction is of great interest in applied machine learning [41,66,7]. Distillation [4] has also been explored as a churn mitigation technique, where researchers aim to transfer knowledge from a baseline model to a new model by regularizing the predictions towards the baseline model [4,184,101,155,79]. This chapter is complementary to this discourse offering a fresh perspective.

Backward Compatibility Model update regression or the decline in performance after a model updates [27] has been a topic of interest in applied ML [151]. Researchers have again explored various mitigation strategies including knowledge distillation [181,179] and probabilistic approaches [162]. This backward compatibility research is closely related to the concept of *forgetting* in machine learning where some component of learning is forgotten [32,132,14,138,63,113].

Uncertainty Quantification Uncertainty in deep learning is most often examined from a Bayesian perspective[108,128]. Many approximate methods for inference have been developed, i.e mean-field variational inference[23,53] and MC Dropout[58]. Deep ensembles[100] often have comparable performance[131] but result in scalability issues at inference time. Predictive uncertainty methods that require only a single model have also been introduced[110,149,152,11,159,28,109,142,154,96,166,104,161]; in particular, we implement the SNGP method[104] given its widespread use in industry settings.

Underspecification and Reproducibility Reproducibility is an anchor of the scientific process[26,62,156,95,118,136,115,167,145], and has garnered discussion in ML from the lens of robustness[36,51]. Recently, research has explored how both reproducibility and generalization relate to "underspecification" [51] which is related to overparametrization as well[10,117,127]. Our examination of near-optimal models resonates with these studies that explore how the ML pipeline can produce deviating outcomes.

4.3 Framework

Our goal is to evaluate the (in)stability of model outputs under future data updates by studying how pointwise predictions change in response to model perturbations at training time. We begin with a classification task with a dataset of n instances, $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where $x_i = [1, x_{i1}, \ldots, x_{id}] \in \mathcal{X} \subseteq \mathbb{R}^{d+1}$ is the feature vector and $y_i \in \{0, 1\}$ is an outcome of interest. We fit a classifier $h : \mathbb{R}^{d+1} \to \{0, 1\}$ from a hypothesis class $\mathcal{H}$ parametrized by $\theta \in \Theta \subseteq \mathbb{R}^d$, and write $L(\,\cdot\,; \mathcal{D})$ for the *loss function*, for example cross entropy, evaluated on dataset $\mathcal{D}$. Throughout, we let $M(h; S) \in \mathbb{R}_+$ denote the performance of $h \in \mathcal{H}$ over a sample S in regards to a *performance metric* $M(h)$, where we assume lower values of $M(h)$ are better. For instance, when working with accuracy, we measure the *Accuracy error*: $M(h) = 1 - \text{Accuracy}(h)$.

4.3.1 Predictive Churn

Definition 16 (Predictive churn [38]). *The* predictive churn *between two models, h_A and h_B, trained successively on modified training data, is the proportion of examples in a sample $i \in S$ whose prediction differs between the two models:*

$$C(h_A, h_B; S) = \frac{1}{|S|} \sum_{i \in S} 1[h_A(\boldsymbol{x}_i) \neq h_B(\boldsymbol{x}_i)]. \tag{4.1}$$

For simplicity, we use shorthand notation $C(h_A, h_B)$ in place of $C(h_A, h_B; S)$.

Consider the following illustrative example. Classifier h_A has accuracy 90% and classifier h_B has accuracy 91%. In the best case, h_B correctly classifies the same 90% as h_A while correcting additional points, resulting in $C(h_A, h_B) = 1\%$, and h_B strictly improves h_A. In the worst case, h_A correctly classifies the 9% of h_B errors and h_B correctly classifies the 10% of h_A errors, resulting in $C(h_A, h_B) = 19\%$. In practice, S_B differs from S_A with added or dropped features or training examples.

4.3.2 Predictive Multiplicity

Predictive multiplicity is the prevalence of conflicting predictions over near-optimal models [112,176,76] commonly referred to as the *ε-Rashomon* set.

Predictive multiplicity with respect to a baseline: The ε-Rashomon set is defined with respect to a *baseline model* that is obtained in seeking a solution to the empirical risk minimization problem, i.e.,

$$h_0 \in \operatorname*{argmin}_{h \in \mathcal{H}} L(h; \mathcal{D}). \tag{4.2}$$

Let h_0 denote this *baseline* classier.

Definition 17 (ε-Rashomon Set w.r.t. h_0)**.** *Given a performance metric M, a baseline model h_0, and error tolerance $\varepsilon > 0$, the* ε-Rashomon set *is the set of competing classifiers $h \in \mathcal{H}$ with performance,*

$$\mathcal{H}_\varepsilon(h_0) := \{h \in \mathcal{H} : M(h; \mathcal{D}) \leq M(h_0; \mathcal{D}) + \varepsilon\}. \tag{4.3}$$

$M(h; \mathcal{D}) \in \mathbb{R}_+$ denotes the performance of $h \in \mathcal{H}$ over a dataset $\mathcal{D}$ in regards to performance metric, $M(h)$. $M(h)$ is typically chosen as the loss function, $M = L(h; \mathcal{D})$, but can also be defined in terms of a direct measure of accuracy [176].

Predictive Multiplicity without a baseline: Long et al. [105] suggest an alternative definition of predictive multiplicity in the context of a randomized training procedure, $\mathcal{T}_{rand}(\mathcal{D})$, that is not defined with respect to a baseline model. For shorthand notation, we leave implicit in the sequel the dependence of $\mathcal{T}_{rand}$ on the dataset $\mathcal{D}$.

Definition 18 (Empirical ε-Rashomon set)**.** *Given a performance metric M, an error tolerance $\varepsilon > 0$, and m models sampled from $\mathcal{T}_{rand}$, the* Empirical ε-Rashomon set *is the set of classifiers $h \in \mathcal{H}$ with performance metric better than ε:*

$$\hat{\mathcal{R}}^m_\varepsilon(\mathcal{T}_{rand}) := \{h_1, h_2, \cdots h_m \ : \ h_k \overset{\text{iid}}{\sim} \mathcal{T}_{rand}, M(h_k; \mathcal{D}) \leq \varepsilon, \forall k \in [m]\}. \tag{4.4}$$

We can also define a concept of ambiguity for this empirical ε-Rashomon set.

Definition 19 (Empirical ε-Ambiguity)**.** *Given the empirical ε-Rashomon set, $\hat{\mathcal{R}}^m_\varepsilon(\mathcal{T}_{rand})$, and a dataset sample, S, the* empirical ε-ambiguity *of a prediction problem is the proportion of examples $i \in S$ assigned conflicting predictions by a classifier in the ε-Rashomon set:*

$$\alpha_\varepsilon(\hat{\mathcal{R}}^m_\varepsilon) := \frac{1}{|S|} \sum_{i \in S} \max_{h,h' \in \hat{\mathcal{R}}^m_\varepsilon} 1[h(x_i) \neq h'(x_i)]. \tag{4.5}$$

For simplicity, we use the following shorthand notation $\alpha_\varepsilon(\hat{\mathcal{R}}^m_\varepsilon)$ in place of $\alpha_\varepsilon(\hat{\mathcal{R}}^m_\varepsilon, S)$.

4.4 Unstable Sets

Our main contribution is to bring the two notions of predictive inconsistency together, which we begin in this section. In addition to considering instability over a sample, we can consider predictive consistency at the individual level.

If there exists a model within the ε-Rashomon set that changes the prediction of an individual instance, we say that example is ε-Rashomon *unstable* according to Def. (18). Similarly, if the prediction of an individual example is expected to change as a result of the successive training of a model, then we say the example is *churn unstable*. We define the set of unstable points as follows.

Definition 20 (ε-Rashomon Unstable Set)**.** *The* ε-Rashomon unstable set *is the set of points in S_{test} for which their prediction changes over a pair of models within the ε-Rashomon set*

$$S^{\mathcal{R}}_{unstable}(\mathcal{R}_\varepsilon, S_{test}) = \{i \in S_{test} \ : \ h(x_i) \neq h'(x_i) \forall h, h' \in \mathcal{R}_\varepsilon\}$$

Definition 21 (Churn Unstable Set)**.** *The* churn unstable set *is the set of points in S_{test} that change*

over a model update from h_A to h_B, i.e.,

$$S^{\mathcal{C}}_{unstable}(h_A, h_B, S_{test}) = \{i \in S_{test} \ : \ h_A(x_i) \neq h_B(x_i)\}$$

Given a fixed S_{test}, we can compare $S^{\mathcal{R}}_{unstable}$ and $S^{\mathcal{C}}_{unstable}$ to characterize the relationship between predictive multiplicity and predictive churn: what is the intersection between the ε-Rashomon *unstable* set and the *churn unstable* set?

Remark. Prior work tends to compute ambiguity over the training set [112,176,174]. If S_{test} is the train dataset, then ε-Rashomon *unstable* examples are simply those that are ambiguous according to definitions in the previous section. Here, we evaluate unseen test points and whether they are ε-Rashomon *unstable*.

4.5 Anticipating Unstable Points

In this section, we consider how to predict whether a new test example will be prone to being ε-Rashomon or churn unstable. We want to understand whether uncertainty quantification can help in identifying such an example. Bayesian methods, as well as ensemble techniques, are the most widely used uncertainty quantification techniques. The Bayesian framework, in particular, aims to provide a posterior distribution on predictions, from which one can sample to calculate predictive variance.

4.5.1 Spectral-Normalized Neural Gaussian Process

Given that Bayesian approaches can be computationally prohibitive when training neural networks, methods have been proposed for uncertainty estimation that require training only a single deep neural network (DNN). Previous work has identified an important condition for DNN uncertainty estimation is that the classifier is aware of the distance between test examples and training examples.

Specifically, Liu et al. [104] propose *Spectral-normalized Neural Gaussian Process* (SNGP) for leveraging Gaussian processes in support of distance awareness. The Gaussian process is approximated using a Laplace approximation, resulting in a closed-form posterior for computing predictive uncertainty. SNGP improves distance awareness by ensuring that (1) the output layer is distance aware by replacing the dense output layer with a Gaussian process, and (2) the hidden layers are distance preserving by applying spectral normalization on weight matrices.

4.5.2 Predicting Churn Directly

Given a sample and the accompanying unstable set $S^{\mathcal{C}}_{unstable}$, we can train a classifier, to predict whether an example is likely to be in the unstable set. We construct a simple classification task with a dataset of n instances, $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, where $x_i = [1, x_{i1}, \ldots, x_{id}] \in \mathcal{X} \subseteq \mathbb{R}^{d+1}$ is the feature vector and $y_i^c \in \{0, 1\}$ is now the label indicating whether the example churned (i.e. $1[x_i \in S^{\mathcal{C}}_{unstable}]$). For analysis, we adjust the feature vector in the following ways:

1. Train on only the feature vector $x_i = [1, x_{i1}, \ldots, x_{id}] \in \mathcal{X} \subseteq \mathbb{R}^{d+1}$
2. Add predicted probabilities from original classification task ($y_i \in [0, 1]$ in § 4.3) to the feature vector for training
3. Add ambiguity indicator, $y_i^r \in \{0, 1\}$, i.e. $1[x_i \in S^{\mathcal{R}}_{unstable}]$ to the feature vector with predictive probabilities and features for training

We make each of these feature adjustments and train a corresponding logistic regression classifier to predict churn. In principle, adding in the predicted probability can add useful information to predict churn because the DNN model is more complex than the churn prediction model. We can compare accuracy and gauge any improvements to determine if the ambiguity improves the prediction of churn. We can also measure the linear relationship or correlation between variables by ana-

lyzing the Pearson Correlation for each configuration. We are particularly interested in correlation between the different feature configurations and churn.

4.5.3 Arbitrariness Reduction via Ensembling

We adopt and implement an ensemble algorithm to examine whether reducing arbitrariness in general will also reduce churn. Long et al. [105] present a simple ensemble algorithm for arbitrariness reduction and detail theoretical guarantees to show that ambiguity is reduced. The ensembling process involves training each model via $\mathcal{T}_{rand}$, then combining those individual predictions to produce a combined prediction. The set of models that is averaged over is exactly an empirical ε-Rashomon set of models.

Definition 22 (Ensemble Classifier [105])**.** *Given the set of models, $\hat{\mathcal{R}}_\varepsilon^m(\mathcal{T}_{rand})$, and a vector $\lambda \in \Delta_m$, the ensemble classifier is the convex combination $h^\lambda := \sum_{j\in[m]} \lambda_j h_j$*

where h_j is the jth model from $\hat{\mathcal{R}}_\varepsilon^m(\mathcal{T}_{rand})$.

For our analysis, we assume the weights $\lambda \in \Delta_m$ to be the vector $\frac{1}{m}$. See Long et al. [105] for a details on parameter optimization.

To calculate ambiguity, we train multiple ensembled classifiers, then determine whether there is predictive disagreement across these ensembled classifiers. Of course, in the large ensemble limit, the disagreement between ensembles becomes zero. In practice, we use a finite ensemble due to limited computational cost.

4.6 Theoretical Results

In this section, we provide theoretical insights into churn using the ε-Rashomon set perspective. Accompanied proofs are included.

We assume that a practitioner only has access to the initial Model A. In § 4.6.1, we derive an analytical bound on the expected churn between Model A and a prospective Model B using only the properties of their respective Rashomon sets. Practically, this implies that if future models are confined to be with the ε-Rashomon set (with respect to a baseline), then the expected churn will be nicely bounded.

Again, operating under the premise that we only have access to Model A, we analyze whether one model within the ε-Rashomon set might result in less churn compared to another model within the ε-Rashomon set. Specifically, we aim to quantify the expected churn difference between any two models within the ε-Rashomon set. In § 4.6.2, we assume that the ε-Rashomon set is defined with respect to a *baseline model* and derive an expected churn difference that resembles prior bounds on discrepancy (see definition in previous chapters) a metric from predictive multiplicity [112,176]. In § 4.6.3, we operate without a baseline and show that the expected churn difference between two models within the ε-Rashomon set can be negligible. These results underscore that the feasibility of mitigating churn by substituting Model A with an alternative from the ε-Rashomon set depends the methodology used to construct the ε-Rashomon set, particularly the presence of a baseline model.

4.6.1 Expected Churn Between Rashomon Sets $\mathcal{H}_\varepsilon(h_0)$

Consider an ε-Rashomon set with respect to a baseline model, $\mathcal{H}_\varepsilon(h_0)$. Say we have two training datasets $\mathcal{D}_A$ and $\mathcal{D}_B$ where $\mathcal{D}_B$ is an updated version of $\mathcal{D}_A$, and consider $\mathcal{H}_\varepsilon(h_0^A)$ and $\mathcal{H}_\varepsilon(h_0^B)$ respectively (where the baseline is defined according to Eq. (4.2) and Eq. (4.3))

We ask what the maximum difference in churn will be between two models from each ε-Rashomon set; i.e., we want to find the worst case scenario in terms of churn between $\mathcal{H}_\varepsilon(h_0^A)$ and $\mathcal{H}_\varepsilon(h_0^B)$. We begin by restating a bound on churn between two models, making use of smoothed churn alongside β-stability [24] of algorithms defined here.

Definition 23 (β-stability[38]). *Let $f_T(x) \mapsto \mathbf{R}$ be a classifier discriminant function (which can be thresholded to form a classifier) trained on a set T. Let T^i be the same as T except with the ith training sample (x_i, y_i) replaced by another sample. Then, as in[24], training algorithm $f(.)$ is β-stable if:*

$$\forall x, T, T^i : |f_T(x) - f_{T^i}(x)| \leq \beta \tag{4.6}$$

We begin by following Cormier et al. [38] to define *smooth* churn and additional assumptions. These assumptions allow us to rewrite churn in terms of zero-one loss:

$$C(h_A, h_B) = \mathbb{E}_{(X,Y)\sim\mathcal{D}} \left[\ell_{0,1}(h_A(X), Y) - \ell_{0,1}(h_B(X), Y)\right],$$

This requires that the data perturbation (update from $\mathcal{D}_A$ to $\mathcal{D}_B$) does not remove any features, that the training procedure is independent of the ordering of data examples, and that training datasets are sampled i.i.d., which ignores dependency between successive training runs.

Cormier et al. [38] also introduce a relaxation of churn called *smooth churn*, which is parametrerized by $\gamma > 0$, and defined as

$$C_\gamma(h_A, h_B) = \mathbb{E}_{(X,Y)\sim\mathcal{D}} \left[\ell_\gamma(f_A(X), Y) - \ell_\gamma(f_B(X), Y)\right],$$

where $f.(X) \in [0, 1]$ is a score that is thresholded to produce the classification $h.(X)$, and ℓ_γ

is defined as

$$\ell_\gamma(f(X), Y) = \begin{cases} 1, & \text{if } f(X)Y < 0, \\ 1 - \frac{f(X)Y}{\gamma}, & \text{if } 0 \leq f(X)Y \leq \gamma, \\ 0, & \text{otherwise.} \end{cases}$$

where $Y \in \{0, 1\}$ here.

Here, γ acts like a confidence threshold. We can use smoothed churn alongside the β-stability [24] (see Definition 23) of algorithms following [38] to derive the bound on expected churn between models within an ε-Rashomon set.

Theorem 1 (Expected Churn between Rashomon Sets)**.** *Assume a training algorithm that is β-stable. Given two ε-Rashomon sets defined with respect to the baseline models, $\mathcal{H}_\varepsilon(h_0^A)$ and $\mathcal{H}_\varepsilon(h_0^B)$, the smooth churn between any pair of models within the two ε-Rashomon sets: $h'_A \in \mathcal{H}_\varepsilon(h_0^A)$ and $h'_B \in \mathcal{H}_\varepsilon(h_0^B)$ is bounded as follows:*

$$\mathbb{E}_{\mathcal{D}_A, \mathcal{D}_B \sim \mathcal{D}^m}[C_\gamma(h'_A, h'_B)] \leq \frac{\beta\sqrt{\pi n}}{\gamma} + 2\varepsilon. \tag{4.7}$$

This holds assuming all models h are trained with randomized algorithms which are also β-stable (Def. 23).

Below, we provide the proof.

Proof of Theorem 1. We first state the results from Cormier et al. [38].

Theorem 2 (Bound on Expected Churn [38])**.** *Assuming a training algorithm that is β-stable, given training datasets $\mathcal{D}_A$ and $\mathcal{D}_B$, sampled i.i.d. from $\mathcal{D}^n$ where two classifiers h_A and h_B are trained on*

$\mathcal{D}_A$ and $\mathcal{D}_B$ respectively, the expected smooth churn obeys:

$$\mathbb{E}_{\mathcal{D}_A,\mathcal{D}_B\sim\mathcal{D}^n}\left[C_\gamma(h_A,h_B)\right] \leq \frac{\beta\sqrt{\pi n}}{\gamma}. \tag{4.8}$$

From Theorem 2, the smooth churn between the two baseline models is bounded by:

$$\mathbb{E}_{\mathcal{D}_A,\mathcal{D}_B\sim\mathcal{D}^m}[C_\gamma(h_0^A,h_0^B)] \leq \frac{\beta\sqrt{\pi n}}{\gamma}.$$

The churn between any two models within the ε-Rashomon sets, $\mathcal{H}_\varepsilon(h_0^A)$ and $\mathcal{H}_\varepsilon(h_0^B)$, is bounded by this constant plus a new 2ε term. To show this, we apply the triangle inequality and Lemma 2, working with any pair of models, $h'_A \in \mathcal{H}_\varepsilon(h_0^A)$ and $h'_B \in \mathcal{H}_\varepsilon(h_0^B)$:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_A,\mathcal{D}_B\sim\mathcal{D}^m}[C_\gamma(h'_A,h'_B)]\\
&= \mathbb{E}_{(X,Y)\sim\mathcal{D}}\left[\ell_\gamma(h'_A(X),Y) - \ell_\gamma(h'_B,Y)\right]\\
&= \mathbb{E}_{(X,Y)\sim\mathcal{D}}[\ell_\gamma(h'_A(X),Y) + \ell_\gamma(h_0^A(X),Y)\\
&- \ell_\gamma(h_0^A(X),Y) + \ell_\gamma(h_0^B(X),Y) - \ell_\gamma(h_0^B(X),Y)\\
&- \ell_\gamma(h'_B,Y)]\\
&= \mathbb{E}_{(X,Y)\sim\mathcal{D}}\left[\ell_\gamma(h'_A(X),Y) - \ell_\gamma(h_0^A(X),Y)\right]\\
&+ \mathbb{E}_{(X,Y)\sim\mathcal{D}}\left[\ell_\gamma(h_0^A(X),Y) - \ell_\gamma(h_0^B(X),Y)\right]\\
&+ \mathbb{E}_{(X,Y)\sim\mathcal{D}}\left[\ell_\gamma(h_0^B(X),Y) - \ell_\gamma(h'_B,Y)\right]\\
&\leq \varepsilon + \frac{\beta\sqrt{\pi n}}{\gamma} + \varepsilon = \frac{\beta\sqrt{\pi n}}{\gamma} + 2\varepsilon,
\end{aligned}$$

where the second and third equalities are algebra. For the inequality, the first and third expectations

follow from the Definition of smooth churn and the middle expectation from Theorem 2. For the final equality, we appeal to Definition 17, with ℓ_γ as the performance metric and ε being the parameter of the Rashomon set. □

4.6.2 Churn for Models within $\mathcal{R}_\varepsilon$

We bound the churn between an optimal baseline model and a model within the ε-Rashomon set. Let $\hat{R}$ denote empirical risk (error) where $\hat{R} := \frac{1}{n}\sum_i 1[h(\boldsymbol{x}_i \neq y_i)]$.

Lemma 1 (Bound on Churn)**.** *The churn between two models h_1 and h_2 is bounded by the sum of the empirical risks of the models:*

$$C(h_1, h_2) \leq \hat{R}(h_1) + \hat{R}(h_2). \tag{4.9}$$

Corollary 3 (Bound on Churn within $\mathcal{R}_\varepsilon$)**.** *Given a baseline model, h_0, and an ε-Rashomon set, $\mathcal{H}_\varepsilon(h_0)$, the churn between h_0 and any classifier in the ε-Rashomon set, $h' \in \mathcal{H}_\varepsilon(h_0)$, is upper bounded by:*

$$C(h_0, h') \leq 2\hat{R}(h_0) + \varepsilon. \tag{4.10}$$

We have recovered a bound on churn that resembles the bound on discrepancy derived in [112] where they show that the discrepancy between the optimal model and a model within the ε-Rashomon set will obey $D_{h_0,\leq}(2;\hat{R})(h_0) + \varepsilon$. Below are the accompanying proofs.

Proof of Lemma 1. This follows from the triangle inequality. For a set $S = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, we denote the predictions as vectors:

$$Y_1 = (h_1(\boldsymbol{x}_1), ..., h_1(\boldsymbol{x}_n)) \in \{0,1\}^n$$

$$Y_2 = (h_2(\boldsymbol{x}_1), ..., h_2(\boldsymbol{x}_n)) \in \{0,1\}^n$$

Let Y denote the ground-truth label,

$$Y = (y_1, ..., y_n) \in \{0,1\}^n.$$

The empirical risk $\hat{R}$ of a classifier can be expressed in terms of the L_1 norm between the predictions and the ground truth:

$$\hat{R}(h_1) = \frac{||Y_1 - Y||_1}{n}, \quad \hat{R}(h_2) = \frac{||Y_2 - Y||_1}{n}$$

Similarly, we write churn as the L_1 norm between the predictions of the two models.

$$C(h_1, h_2) = \frac{||Y_1 - Y_2||_1}{n}$$

The triangle inequality results in:

$$||Y_1 - Y_2||_1 \leq ||Y_1 - Y||_1 + ||Y - Y_2||_1$$

Substitution and dividing by n gives

$$C(h_1, h_2) \leq \hat{R}(h_1) + \hat{R}(h_2).$$

□

Proof of Corollary 3. By definition, $\hat{R}(h') \leq \hat{R}(h_0) + \varepsilon$. Following Lemma 1, we have:

$$C(h_0, h') \leq \hat{R}(h_0) + \hat{R}(h') \leq 2\hat{R}(h_0) + \varepsilon. \tag{4.11}$$

□

4.6.3 Expected Churn within $\hat{\mathcal{R}}_{\varepsilon}^{m}(\mathcal{T}_{RAND})$

Consider a randomized training procedure $\mathcal{T}_{rand}(\mathcal{D})$ over a hypothesis class $\mathcal{H}$ and a fixed finite dataset $\mathcal{D}$. Say we derive the empirical ε-Rashomon set, $\hat{\mathcal{R}}_{\varepsilon}^{m}(\mathcal{T}_{rand})$, according to Def. 18. We ask whether there is a model within this empirical ε-Rashomon set that might decrease churn if used as an alternative starting point for the successive training of two models. Said another way, we are interested in whether switching one model out for another within the ε-Rashomon set will impact churn.

Given $\mathcal{T}_{rand}(\mathcal{D})$ is a randomized training procedure, we show there is no difference in expected churn when adopting any two models in $\hat{\mathcal{R}}_{\varepsilon}^{m}(\mathcal{T}_{rand})$ as h_A and h'_A, and considering churn with respect to some other model h_B.

Lemma 2 (Same Expected Churn within $\hat{\mathcal{R}}_{\varepsilon}^{m}(\mathcal{T}_{rand})$)**.** *Assume a randomized training procedure $\mathcal{T}_{rand}(\mathcal{D})$. Fix a training dataset $\mathcal{D}_A$ and an arbitrary model h_B. Let h_A and h'_A be two models induced by $\mathcal{T}_{rand}(\mathcal{D}_A)$. The expected difference in churn between any models h_A and h'_A induced by $\mathcal{T}_{rand}(\mathcal{D}_A)$ is zero*

$$\mathbb{E}_{h_A, h'_A \overset{iid}{\sim} \mathcal{T}_{rand}(\mathcal{D}_A)} \left[C(h_A, h_B) - C(h'_A, h_B) \right] = 0$$

This means that one model sampled from T_{rand} will have the same expected churn as another model sampled from T_{rand}. In essence, we will not reduce churn by replacing the current model with one from the ε-Rashomon set when using the randomized approximation approach. The proof is below.

Proof of Lemma 2. We use linearity of expectation and the assumption that models in $\mathcal{T}_{\mathcal{D}_A}$ are sam-

pled i.i.d. to show that the difference in expectation is 0.

$$
\begin{aligned}
&\mathbb{E}_{h_A, h'_A \overset{iid}{\sim} \mathcal{T}_{rand}(\mathcal{D}_A)} \left[C(h_A, h_B) - C(h'_A, h_B) \right] \\
&= \mathbb{E}_{h_A \overset{iid}{\sim} \mathcal{T}_{rand}(\mathcal{D}_A)} [C(h_A, h_B)] \\
&\quad - \mathbb{E}_{h'_A \overset{iid}{\sim} \mathcal{T}_{rand}(\mathcal{D}_A)} [C(h'_A, h_B))] \\
&= \mathbb{E}_{h_A \overset{iid}{\sim} \mathcal{T}_{rand}(\mathcal{D}_A)} [\mathbb{E}_{(X,Y)\sim\mathcal{D}} [\ell_{0,1}(h_A(X), Y) - \ell_{0,1}(h_B)(X), Y)]] \\
&\quad - \mathbb{E}_{h'_A \overset{iid}{\sim} \mathcal{T}_{rand}(\mathcal{D}_A)} [\mathbb{E}_{(X,Y)\sim\mathcal{D}} [\ell_{0,1}(h'_A(X), Y) - \ell_{0,1}(h_B)(X), Y)]] \\
&= \mathbb{E}_{h_A, h'_A \overset{iid}{\sim} \mathcal{T}_{rand}(\mathcal{D}_A)} [\mathbb{E}_{(X,Y)\sim\mathcal{D}} [\ell_{0,1}(h_A(X), Y) - \ell_{0,1}(h'_A(X), Y)]] \\
&= \mathbb{E}_{h_A \overset{iid}{\sim} \mathcal{T}_{rand}(\mathcal{D}_A)} [\mathbb{E}_{(X,Y)\sim\mathcal{D}} [\ell_{0,1}(h_A(X), Y)]] - \mathbb{E}_{h'_A \overset{iid}{\sim} \mathcal{T}_{rand}(\mathcal{D}_A)} [\mathbb{E}_{(X,Y)\sim\mathcal{D}} [\ell_{0,1}(h'_A(X), Y)]] \\
&= 0.
\end{aligned}
$$

□

4.7 Experiments

In this section, we present experiments on real-world datasets in domains where predictive instability is particularly high-stakes (i.e. lending and housing).

4.7.1 Setup

Datasets. We consider datasets with varying sample size, number of features, and class imbalance; summary statistics for each dataset are in table 4.1. * As we show below, models trained and tested on these datasets exhibit notable variation in predictive inconsistency.

*Notice that the HMDA dataset is an order of magnitude larger than the others ($n = 244,107$).

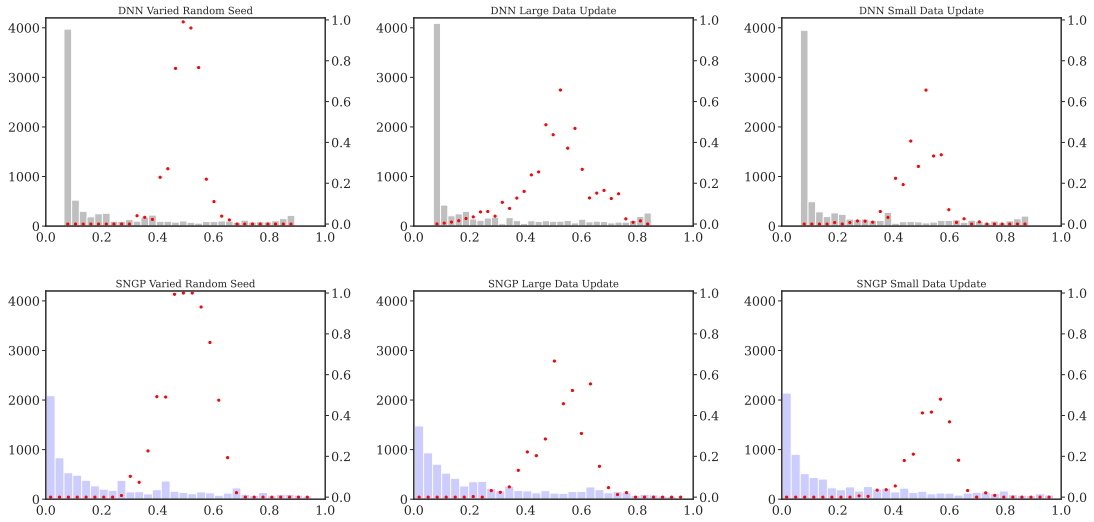


Figure 4.1: Predicted probability distributions for the Adult Dataset. We plot a histogram of predicted probability distribution in grey with the left y-axis ($0 - 4000$ are counts) and a scatter plot of the proportion of flip counts for each bin aligned with the right y-axis ($0 - 1$ is a proportion). The x-axis is the predicted probability. By overlapping the plots, we gain a comprehensive view of the model's confidence in its predictions (via the histogram) and the areas where the model predictions are most prone to change (scatter plot of flips). Notice that the scale is different between the histogram and the flip counts. The top row corresponds to the DNN experiments and the bottom row are the UA-DNN experiments. Each column represents an experiment. From the left, we show results for predictive multiplicity, large dataset update, and small dataset update.

Metrics. We measure predictive inconsistency by computing the measures detailed in § 4.3. In terms of predictive multiplicity, we compute the empirical ε-Rashomon set and report ε-ambiguity over a test sample according to either Eq (4.5). When training sets of models, we use multiple arrays of random seeds $\{0.0, 1.0, 109, 10, 1234\}$, $\{3666, 2299, 2724, 1262, 4220\}$, $\{3971, 9444, 1375, 7351, 2083\}$, $\{1429, 2281, 2189, 9376, 2261\}$ and $\{1881, 2273, 9509, 6707, 4412\}$. For varying random initialisations, we repeat experiments across these arrays. As in § E.2 in Long et al. [105], we can set ε in the definition of the empirical ε-Rashomon set to the worst value of the performance metric over the generated trained models. As a result, the experiments on predictive multiplicity reported in this chapter do not need to be explicitly parametrized by ε. In terms of predictive churn, we report over a test sample according to

Eq. (4.1).

Churn Regimes. We compute *predictive churn* Eq. (4.1) for different types of successive training updates according to literature on predictive churn [38]. First, we imitate a large dataset update by comparing Model B (h_B) trained on the full dataset to Model A (h_A) trained on a random sample of half the dataset. Second, we imitate a small dataset update by comparing Model B (h_B) trained on the full dataset to Model A (h_A) trained on a random sample of 95% the dataset – i.e. 5% of examples have been dropped or added between the two models. These two updates are similar but represent two different regimes (see Giordano et al. [64]). †

Model Classes. We consider two classes of deep neural networks (DNNs). We train a standard neural network made up of 1 or more layers and refer to this as DNN. We also train a DNN that incorporates an uncertainty awareness technique and refer to this as UA-DNN. For this demonstration, we implement the SNGP technique described in § 4.5.1 to train the uncertainty aware model, UA-DNN. To ensure the models are well calibrated, we tune the parameters within the SNGP technique and apply Platt scaling for the standard DNN.

Experiment Details.

Models All models use a shallow neural network with 1 or more fully connected layers. There is 1 hidden layer with 279 units, learning rate of 0.0000579, dropout rate of 0.0923 and batch normalization is enabled. All training is conducted in TensorFlow with a batch size of 128. For churn experiments, we use the first random seed in the array as the default seed and repeat experiments across these values. We run on a single CPU with 50GB RAM.

The SNGP training process follows the standard DNN learning pipeline, with the updated Gaussian process and spectral normalization outputting predictive logits and posterior covariance. The steps for SNGP prediction are as follows. For a test example, the model posterior

† For instance, leave-one-out jackknife is a small data perturbation, whereas bootstrap is a large data perturbation; see papers on infinitesimal jackknife i.e. Giordano et al. [64].

Dataset Name	Outcome Variable	n	d	Class Imbalance
Adult[92]	person income over $50,000	16,256	28	0.31
HMDA[36]	loan granted	244,107	18	3.3
Credit[182]	customer default on loan	30,000	23	3.50
Mammo[52]	mammogram shows breast cancer	961	12	0.86

Table 4.1: Datasets used in the experiments. For each dataset, we report n, d and the class imbalance ratio of a model on test data.

Dataset	Model	Predictive Multiplicity (Empirical ε-Ambiguity)	AUC	Predictive Churn (Large Data Update)	AUC	Predictive Churn (Small Data Update)	AUC
Adult	DNN	0.047 ± 0.003	0.89 ± 0.010	0.058 ± 0.004	0.89 ± 0.009	0.028 ± 0.004	0.89 ± 0.01
Credit	DNN	0.053 ± 0.004	0.76 ± 0.01	0.050 ± 0.004	0.76 ± 0.009	0.029 ± 0.004	0.76 ± 0.01
HMDA	DNN	0.021 ± 0.004	0.89 ± 0.011	0.042 ± 0.004	0.89 ± 0.009	0.007 ± 0.004	0.89 ± 0.01
mammo	DNN	0.007 ± 0.0018	0.83 ± 0.001	0.027 ± 0.024	0.85 ± 0.007	0.014 ± 0.017	0.83 ± 0.004
Adult	UA-DNN	0.12 ± 0.010	0.87 ± 0.015	0.074 ± 0.011	0.84 ± 0.012	0.041 ± 0.008	0.87 ± 0.016
Credit	UA-DNN	0.10 ± 0.010	0.76 ± 0.015	0.067 ± 0.012	0.76 ± 0.012	0.05 ± 0.008	0.76 ± 0.016
HMDA	UA-DNN	0.14 ± 0.010	0.87 ± 0.015	0.12 ± 0.011	0.84 ± 0.013	0.06 ± 0.008	0.87 ± 0.016
mammo	UA-DNN	0.047 ± 0.013	0.82 ± 0.001	0.041 ± 0.019	0.83 ± 0.005	0.025 ± 0.020	0.83 ± 0.004

Table 4.2: This table shows that predictions are more sensitive to model perturbations (multiplicity) and an uncertainty-aware (UA) model can exhibit higher ambiguity compared to a standard DNN. We compare predictive multiplicity and predictive churn across datasets and model specifications. Over a held out sample, we compute empirical ε-ambiguity, as well as churn induced by a large or small data update. We also show the range of AUC over runs for each.

mean and covariance are used to compute the predictive distribution. Specifically, we approximate the posterior predictive probability, $E(p(x))$, using the mean-field method $E(p(x)) \sim \text{softmax}\left(\text{logit}(x)/\sqrt{1+\lambda * \sigma^2(x)}\right)$, where $\sigma^2(x)$ is the SNGP variance and λ is a hyperparameter, tuned for optimal model calibration (in deep learning, this is known as temperature scaling[67]).

Dataset	Model	Predictive Multiplicity (Empirical ε-Ambiguity)	AUC	Predictive Churn (Large Data Update)	AUC	Predictive Churn (Small Data Update)	AUC
Adult	DNN	0.004 ± 0.001	0.89 ± 0.001	0.002 ± 0.006	0.89 ± 0.001	0.003 ± 0.001	0.89 ± 0.001
Credit	DNN	0.005 ± 0.0004	0.76 ± 0.002	0.003 ± 0.0001	0.76 ± 0.004	0.0028 ± 0.0004	0.76 ± 0.002
HMDA	DNN	0.005 ± 0.001	0.90 ± 0.0003	0.004 ± 0.001	0.90 ± 0.0004	0.003 ± 0.001	0.90 ± 0.0003
mammo	DNN	0.004 ± 0.003	0.86 ± 0.003	0.004 ± 0.003	0.85 ± 0.009	0.002 ± 0.002	0.85 ± 0.01
Adult	UA-DNN	0.0 ± 0.0	0.89 ± 0.002	0.028 ± 0.0001	0.87 ± 0.002	0.019 ± 0.002	0.88 ± 0.003
Credit	UA-DNN	0.0 ± 0.0	0.75 ± 0.004	0.035 ± 0.003	0.75 ± 0.006	0.020 ± 0.002	0.75 ± 0.003
HMDA	UA-DNN	0.0 ± 0.0	0.90 ± 0.001	0.046 ± 0.002	0.90 ± 0.0001	0.041 ± 0.002	0.90 ± 0.0002
mammo	UA-DNN	0.0 ± 0.0	0.84 ± 0.003	0.02 ± 0.009	0.83 ± 0.010	0.005 ± 0.006	0.84 ± 0.008

Table 4.3: Ensemble Results. We compare predictive multiplicity and predictive churn across datasets and model specifications. Over a held out sample, we compute empirical ambiguity, as well as churn induced by a large or small data update. We also show the range of AUC over runs for each.

4.7.2 Results

Predictive Multiplicity vs Predictive Churn. We investigate whether the severity of predictive churn between Model A and Model B is captured by predictive multiplicity analysis on only Model A. Findings for the Standard DNN and UA-DNN are shown in Table 4.2. Notably, we see that model performance, as measured by AUC, is largely uniform across the table: random seed/data perturbations (columns) do not seem to affect overall predictive performance whereas AUC of the UA-DNN is less than or equal to that of DNN. Thus, under standard criteria for evaluating ML models, differences in prediction across these models could be considered to be "arbitrary", with the corresponding implications for fairness discussed above.

We highlight several notable patterns. First, although they are measured on similar scales, predictive multiplicity as measured via ambiguity tends to be larger than predictive churn. Thus, in the settings that we study, predictions appear to be broadly more sensitive to model perturbations than to data updates. But only by a small amount. In terms of gauging the severity of potential predictive arbitrariness, analyzing ambiguity may help to anticipate predictive churn.

Second, within model specifications (DNN or UA-DNN), predictive multiplicity and predictive churn measurements generally align. Specifically, when a model exhibits high predictive multiplicity on one dataset relative to others, it also exhibits high predictive churn (across both churn

regimes) relative to other datasets. Thus, for a given model, it is possible that the same properties of the dataset drive predictive multiplicity and predictive churn.

However, interestingly, between the DNN and UA-DNN specifications, we see that different datasets exhibit high prediction instability. For example, while DNN exhibits high(er) predictive multiplicity on Credit, the UA-DNN exhibits higher predictive multiplicity on HMDA but relatively lower on Credit. This highlights that arbitrariness in predictions is driven by an interaction between the dataset and the model specification, not by the data itself; echoing predictive arbitrariness studies from algorithmic fairness [36]. Importantly, this also highlights that a particular model specification may not be a general solution for mitigating arbitrariness across all settings.

Comparison of Unstable Sets. We examine whether examples that are unstable over the update between Model A and Model B are included in those flagged as unstable when only using the ε-Rashomon set of Model A. We study the broad patterns highlighted above in more detail by comparing the ε-Rashomon unstable set to the churn unstable set for a given test sample S_{test}. For a given dataset, we take a heldout test sample and compute $S^{\mathcal{R}}_{unstable}(S_{test})$ and $S^{\mathcal{C}}_{unstable}(S_{test})$ described in § 4.4. Given that $\#\{S^{\mathcal{R}}_{unstable}(S_{test})\}$ tends to be greater than $\#\{S^{\mathcal{C}}_{unstable}(S_{test})\}$, we calculate what proportion of test examples in $\#\{S^{\mathcal{C}}_{unstable}(S_{test})\}$ are contained in $\#\{S^{\mathcal{R}}_{unstable}(S_{test})\}$ and report this common arbitrariness.

For instance, if all the examples in S_{test} that churn are contained in the ε-Rashomon unstable set, then the common arbitrariness would be 100%. If none of the examples in S_{test} that churn are contained in the ε-Rashomon unstable set then the common arbitrariness would be 0%. Results are show in Table 4.4. As expected, for the small data updates, the common arbitrariness is much higher than compared to the large data update. Comparing model classes, the UA-DNN for small dataset updates seems to recover the largest overlap ranging between 81% to 91% across datasets.

Predicted Probabilities and Unstable Examples. Finally, we examine how predicted probabilities relate to which points are identified as unstable. With the ε-Rashomon unstable set and the churn unstable sets over a given test sample, we visualize the number of unstable examples alongside the full predicted probability distribution in Figure 4.1. First, we plot a histogram of the predicted probabilities for the test sample. Then, for each bin of the histogram, we compute the counts of the unstable (flipped) examples within that bin. Namely, the number of unstable (flipped) examples in a bin divided by the total number of predictions in that bin. This highlights where the model's predictions are most unstable or uncertain as indicated by a higher proportion of unstable points. For additional datasets, we plot a histograms of predicted probability distributions in Figure 4.3 and Figure 4.4.

We see that predicted probabilities of flipped examples (red points) are similarly concentrated in the middle of the unit interval comparing DNN to UA-DNN, which is somewhat surprising given the explicit consideration of uncertainty in UA-DNN. But one side effect of this consideration is that small perturbations may send UA-DNN predictions across the default decision boundary, which could explain the generally higher rates of arbitrariness in Table 4.2, especially under the predictive multiplicity perspective.

The findings suggests that UA-DNN models can provide useful indications of which examples are more at risk of being unstable under perturbations of the UA-DNN model, as a result of both predictive multiplicity or churn. Hence, the results show that model specification may not be the driving factor here. The predicted probabilities around the threshold (0.5) are more likely to be unstable. Therefore, the important difference in model type seems to be calibration.

Predicting Churn. As described in § 4.5.2, we can train a classifier to predict churn to examine correlation between ambiguity and predictive churn. First, we examine the correlation between variables by analyzing the Pearson Correlation between the features, predicted probabilities, ambi-

Dataset	Model	Predictive Churn (Large Data Update)	Predictive Churn (Small Data Update)
Adult	DNN	0.58	0.73
Credit	DNN	0.47	0.85
HDMA	DNN	0.68	0.78
mammo	DNN	0.20	0.50
Adult	UA-DNN	0.64	0.91
Credit	UA-DNN	0.67	0.81
HDMA	UA-DNN	0.44	0.81
mammo	UA-DNN	0.73	1.0

Table 4.4: This table shows the ε-Rashomon unstable set tends to contain many of the examples within the churn unstable set. We report common flipped examples across different experiments i.e. the proportion of churned examples that are included in the ε-Rashomon unstable set.

guity indicator and churn indicator. We are most interested in correlation between ambiguity and churn. In Figure 4.2, there is not much correlation between ambiguity and churn for the mammo and adult datasets (top left and right). But there does seem to be a negative correlation for the hmda and credit datasets (bottom left and right). Second, in regard to the classifier to predict churn for different feature configurations, the results are inconclusive, with little effect of different feature configurations on predictive accuracy.

Ambiguity and Churn for Ensemble Classifiers. Given that ensembling is a technique used to decrease ambiguity[21,105], we compute ambiguity and churn for ensemble classifiers showing results in Table 4.3. Notably, the ambiguity for the uncertainty aware model is zero across datasets. And churn has decreased significantly as well. These results support the intuition that arbitrariness reduction is related to churn reduction.

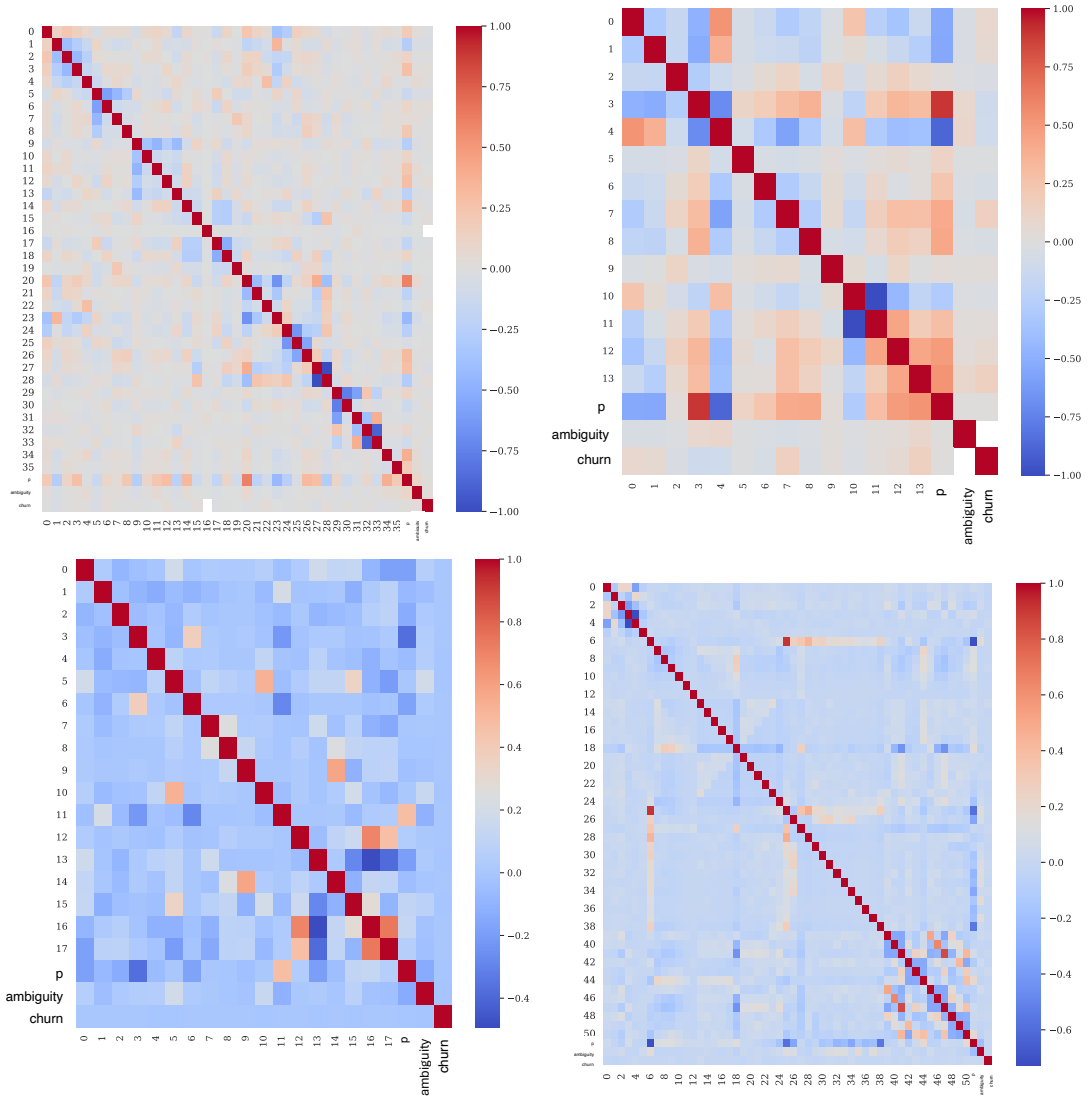


Figure 4.2: Pearson correlation between features, predicted probabilities (p), ambiguity indiciator and churn indicator. Top left is *adult*, top right is *mammo*, bottom left is *hmda*, bottom right is *credit*. Results shown for DNN model.

4.8 Implications

Our findings reveal that analyzing predictive multiplicity is a useful way to anticipate predictive churn over time. We can consider the set of *prospective models* around the selected deployed model

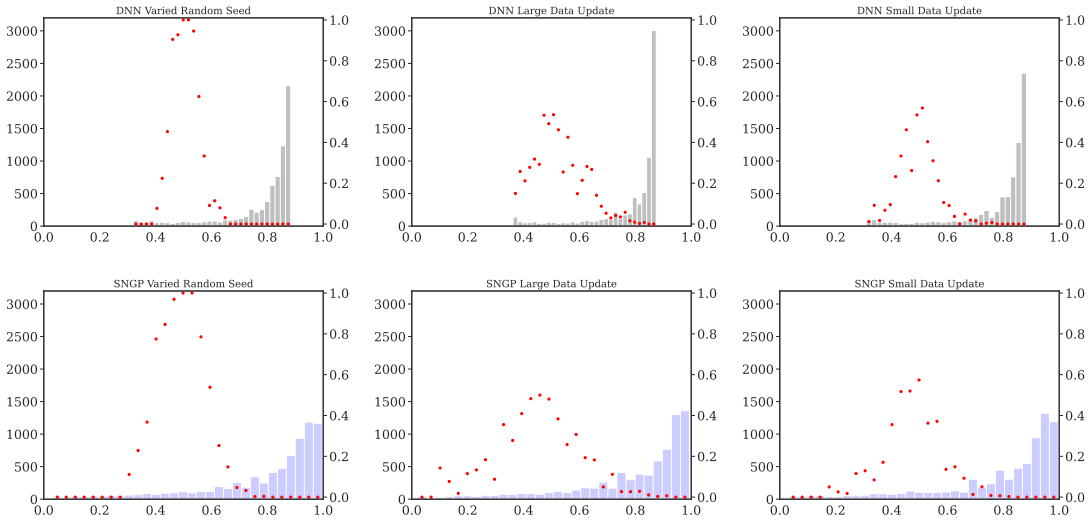


Figure 4.3: Predicted probability distributions for Credit Dataset. We plot a histogram of predicted probability distribution in grey with the left y-axis ($0 - 4000$ are counts) and a scatter plot of the proportion of flip counts for each bin aligned with the right y-axis ($0 - 1$ is a proportion). The x-axis is the predicted probability.

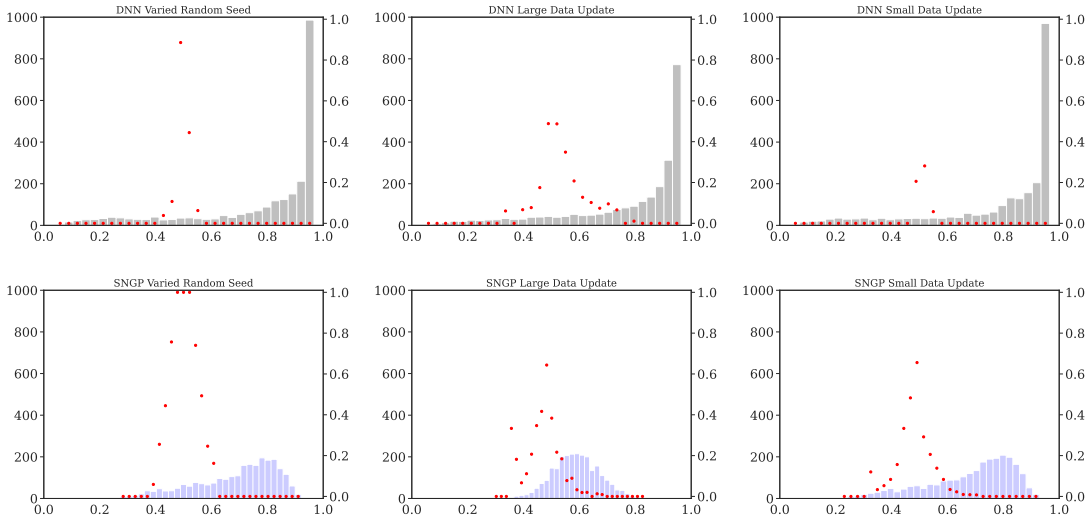


Figure 4.4: Predicted probability distributions for HDMA Dataset. We plot a histogram of predicted probability distribution in grey with the left y-axis ($0 - 4000$ are counts) and a scatter plot of the proportion of flip counts for each bin aligned with the right y-axis ($0 - 1$ is a proportion). The x-axis is the predicted probability.

and draw conclusions about anticipated predictive churn. Given that research in predictive multiplicity has largely focused on how to measure its severity and methods to train the ε-Rashomon set, the present study demonstrates how predictive multiplicity can help assess an important notion of predictive instability (churn).

To combine predictive multiplicity and churn, a practitioner could conduct one analysis after the other. For choosing a better starting point while anticipating model updates, we can begin with a predictive multiplicity analysis following by a predictive churn analysis. Say for instance, we have a model A that we are considering for deployment. We can ask if there might exist a model within the ε-Rashomon set for which the anticipated churn is likely less than that of model A. To do this, we can train the ε-Rashomon set with model A as a baseline then evaluate changes in the churn unstable set for each model within the Rashomon set. We can also train the ε-Rashomon set without assuming a baseline and choose the model that might minimize expected churn from that.

Previous studies have examined various churn reduction methods [38,79]. It will be interesting in future work to examine whether known churn reduction methods (e.g., distillation and constrained weight optimization) might improve predictive multiplicity. To do this, we would analyze predictive multiplicity over a standard training procedure then, make improvements to said training procedure that for churn reduction and analyze predictive multiplicity over this improved training procedure. Similar to our empirical demonstrations, you can then take a fixed test set and compare the ε-Rashomon unstable set against the churn unstable set. Ultimately, this would provide insight into whether training procedures that are more robust to churn are also more robust to predictive multiplicity. And, in line with bridging between uncertainty quantification and fairness as arbitrariness, future work can also explore additional methods from reliable deep learning i.e [161].

4.9 Concluding Remarks

Reducing arbitrariness in machine learning is critical for machine learning credibility and reproducibility. This goal aligns with efforts to address arbitrariness as as a form of unfairness. In particular, fairness researchers underline the challenge in justifying the use of predictions to inform decision making when there exists equally good models that might change individual outcomes[22]. We advocate for connecting the fairness/safety perspective to research on reliable and robust learning. This study is an initial step in that direction.

5

Conclusion

In this dissertation, I have presented research on predictive multiplicity in machine learning. Methods to measure predictive multiplicity in probabilistic classification are detailed in Chapter 2. Compared to previous work [112], our methods allow for flexibility in choosing near-optimal metric and deviation threshold. Our results show that ambiguity and discrepancy vary considerably based on the defining near-optimal in terms of AUC or loss. The analysis of viable prediction ranges shows that models in the Rashomon set can assign risk estimates that vary substantially. Empirical results

on real-world datasets echo the synthetic dataset analysis on dataset characteristics the lead to more or less ambiguity. Namely, empirical results reveal a relationship between ambiguity and individual *uniqueness* (number of duplicates), *class* imbalance, and *baseline risk estimate*. Numerical experiments show that ambiguity increases with dataset separability, examples that are outliers yet far from the discriminant boundary are less prone to ambiguity, and the ambiguity for a minority group is much larger than that for the majority group (given a majority-minority structure).

The notion of "multi-target multiplicity" is introduced and outlined in Chapter 3 along with an extension of standard measures of predictive multiplicity to the resource constrained setting. The framework shows that when considering multiple target variable options, practitioners can develop index models that address fairness concerns (selection rate disparities) by re-weighting and combining predictions for each target. Experiment results show that the framework is effective for narrowing racial disparities in selection rates in an example healthcare allocation task (by Obermeyer et al. [129]) where the goal is to choose patients for a high-risk coordinated care management program.

In Chapter 4, predictive multiplicity is leveraged in support of examining predictive churn. To explore the relationship between these two concepts, I have considered whether individual predictions that are unstable under model perturbations (multiplicity) are also those that are unstable under dataset perturbations (churn). First, results show that although they are measured on similar scales, predictive multiplicity as measured via ambiguity tends to be larger than predictive churn. Thus, in the settings that we study, predictions appear to be broadly more sensitive to model perturbations than to data updates by a small amount on average. In terms of overlap between points that are churn unstable versus Rashomon unstable, the UA-DNN for small dataset updates seems to recover the largest overlap ranging between 81% to 91% across datasets. As for predicted probabilities, the unstable points are concentrated near the middle (50% probability) across experiments. Finally, results show that ensemble classifiers decrease both ambiguity and churn.

For future directions, communicating predictive multiplicity effectively is an important area of

study. This could involve perspectives from data visualization research as well as human-computer interaction research [86,170]. At present, the Rashomon set and the problem of predictive multiplicity tend to receive attention from researchers motivated by algorithmic fairness, transparency and safety. There is a growing need to ensure that these questions of predictive arbitrariness are being connected to ongoing research in applied machine learning that may not share the same motivations. For instance, there are discussions about arbitrariness in recommender systems [130] and other areas of research that are mostly separate from the model multiplicity discourse. It will be important and interesting to draw these connections more directly. Particularly, with respect to mitigation and response strategies, there is more work to be done to investigate the relationship between concepts on predictive arbitrariness like predictive multiplicity and uncertainty quantification research. Along these lines, exploring more state of the art reliable deep learning methods for predictive arbitrariness would be a clear next step. Finally, it will be critical to continue to develop measures that promote transparency on the part of technical researchers. This is especially true with the recent rise of large language models. It will also be interesting to refine or adapt multiplicity concepts to the generative setting.

References

[1] Ali, J., Lahoti, P., & Gummadi, K. P. (2021). *Accounting for Model Uncertainty in Algorithmic Discrimination*. Number 1. Association for Computing Machinery.

[2] Ananny, M. & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973–989.

[3] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias — ProPublica.

[4] Anil, R., Pereyra, G., Passos, A. T., Ormandi, R., Dahl, G., & Hinton, G. (2018). Large scale distributed neural network training through online distillation. In *ICLR*.

[5] Attigeri, G. V., Pai, M. M., & Pai, R. M. (2017). Credit risk assessment using machine learning algorithms. *Advanced Science Letters*, 23(4), 3649–3653.

[6] Austin, J., Ocker, R., & Bhati, A. (2010). Kentucky Pretrial Risk Assessment Instrument Validation. *The JFA Institute*, 5.

[7] Bahri, D. & Jiang, H. (2021). Locally adaptive label smoothing for predictive churn.

[8] Barocas, S. & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104, 671.

[9] Bekhet, H. A. & Eletter, S. F. K. (2014). Credit risk assessment model for Jordanian commercial banks: Neural scoring approach. *Review of Development Finance*, 4(1), 20–28.

[10] Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32), 15849–15854.

[11] Bendale, A. & Boult, T. E. (2016). Towards open set deep networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1563–1572). Los Alamitos, CA, USA: IEEE Computer Society.

[12] Bertsimas, D. & King, A. (2017). Logistic regression: From art to science. *Statistical Science*, (pp. 367–384).

[13] Bertsimas, D., King, A., Mazumder, R., et al. (2016). Best subset selection via a modern optimization lens. *Annals of statistics*, 44(2), 813–852.

[14] Biesialska, M., Biesialska, K., & Costa-jussà, M. R. (2020). Continual lifelong learning in natural language processing: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 6523–6541). Barcelona, Spain (Online): International Committee on Computational Linguistics.

[15] Binns, R. (2018). Algorithmic accountability and public reason. *Philosophy & Technology*, 31(4), 543–556.

[16] Binns, R. (2020). On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20 (pp. 514–524). New York, NY, USA: Association for Computing Machinery.

[17] Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). "it's reducing a human being to a percentage": Perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18: ACM.

[18] Black, E., Elzayn, H., Chouldechova, A., Goldin, J., & Ho, D. (2022a). Algorithmic fairness and vertical equity: Income fairness with irs tax audit models. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1479–1503).

[19] Black, E. & Fredrikson, M. (2021). Leave-one-out unfairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21 (pp. 285–295). New York, NY, USA: Association for Computing Machinery.

[20] Black, E., Koepke, J. L., Kim, P., Barocas, S., & Hsu, M. (2024). Less discriminatory algorithms. *Georgetown Law Journal*, 113(1). Washington University in St. Louis Legal Studies Research Paper Forthcoming. Available at SSRN: https://ssrn.com/abstract=4590481 or http://dx.doi.org/10.2139/ssrn.4590481.

[21] Black, E., Leino, K., & Fredrikson, M. (2021). Selective Ensembles for Consistent Predictions. (NeurIPS), 1–24.

[22] Black, E., Raghavan, M., & Barocas, S. (2022b). Model multiplicity: Opportunities, concerns, and solutions. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 850–863).

[23] Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15 (pp. 1613–1622).: JMLR.org.

[24] Bousquet, O. & Elisseeff, A. (2000). Algorithmic stability and generalization performance. In T. Leen, T. Dietterich, & V. Tresp (Eds.), *Advances in Neural Information Processing Systems*, volume 13: MIT Press.

[25] Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3), 199–215.

[26] Buckheit, J. B. & Donoho, D. L. (1995). Wavelab and reproducible research. In *Wavelets and Statistics*.

[27] Cai, D., Mansimov, E., Lai, Y.-A., Su, Y., Shu, L., & Zhang, Y. (2022). Measuring and reducing model update regression in structured prediction for NLP. In A. H. Oh, A. Agarwal, D. Belgrave, & K. Cho (Eds.), *Advances in Neural Information Processing Systems*.

[28] Calandra, R., Peters, J., Rasmussen, C. E., & Deisenroth, M. P. (2016). Manifold gaussian processes for regression.

[29] Calders, T., Kamiran, F., & Pechenizkiy, M. (2009). Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops* (pp. 13–18).

[30] Chatfield, C. (1995). Model Uncertainty, Data Mining and Statistical Inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 158(3), 419.

[31] Chen, I. Y., Joshi, S., Ghassemi, M., & Ranganath, R. (2021). Probabilistic machine learning for healthcare. *Annual Review of Biomedical Data Science*, 4, 393–415.

[32] Chen, Z., Liu, B., Brachman, R., Stone, P., & Rossi, F. (2018). *Lifelong Machine Learning*. Morgan & Claypool Publishers, 2nd edition.

[33] Choi, D., Shallue, C. J., Nado, Z., Lee, J., Maddison, C. J., & Dahl, G. E. (2019). On empirical comparisons of optimizers for deep learning. *CoRR*, abs/1910.05446.

[34] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2), 153–163. PMID: 28632438.

[35] Christin, A., Rosenblat, A., & Boyd, D. (2015). Courts and predictive algorithms. *Data & civil rights: A new era of policing and justice*, 13.

[36] Cooper, A. F., Lee, K., Choksi, M. Z., Barocas, S., Sa, C. D., Grimmelmann, J., Kleinberg, J., Sen, S., & Zhang, B. (2024). Arbitrariness and social prediction: The confounding role of variance in fair classification.

[37] Cooper, A. F., Lu, Y., Forde, J., & De Sa, C. M. (2021). Hyperparameter Optimization Is Deceiving Us, and How to Stop It. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang,

& J. W. Vaughan (Eds.), *Advances in Neural Information Processing Systems*, volume 34 (pp. 3081–3095).: Curran Associates, Inc.

[38] Cormier, Q., Milani Fard, M., Canini, K., & Gupta, M. R. (2016). Launch and iterate: Reducing prediction churn. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, volume 29: Curran Associates, Inc.

[39] Coston, A., Kawakami, A., Zhu, H., Holstein, K., & Heidari, H. (2022). A validity perspective on evaluating the justified use of data-driven decision-making algorithms. *arXiv preprint arXiv:2206.14983*.

[40] Coston, A., Rambachan, A., & Chouldechova, A. (2021). Characterizing Fairness Over the Set of Good Models Under Selective Labels.

[41] Cotter, A., Jiang, H., Wang, S., Narayan, T., You, S., Sridharan, K., & Gupta, M. R. (2019). Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research*.

[42] Creel, K. & Hellman, D. (2022). The algorithmic leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision-making systems. *Canadian Journal of Philosophy*, 52(1), 26–43.

[43] De-Arteaga, M., Feuerriegel, S., & Saar-Tsechansky, M. (2022). Algorithmic fairness in business analytics: Directions for research and practice. *Production and Operations Management*, 31(10), 3749–3770.

[44] Diakopoulos, N. (2015). Algorithmic accountability. *Digital Journalism*, 3(3), 398–415.

[45] Diakopoulos, N. & Koliska, M. (2017). Algorithmic transparency in the news media. *Digital Journalism*, 5(7), 809–828.

[46] Diamond, S. & Boyd, S. (2016). CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17, 1–5.

[47] Dong, J. & Rudin, C. (2020). Exploring the cloud of variable importance for the set of all good models. *Nature Machine Intelligence*, 2(12), 810–824.

[48] Donnelly, J., Katta, S., Rudin, C., & Browne, E. P. (2023). The rashomon importance distribution: Getting rid of unstable, single model-based variable importance. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[49] Dusenberry, M. W., Tran, D., Choi, E., Kemp, J., Nixon, J., Jerfel, G., Heller, K., & Dai, A. M. (2020). Analyzing the role of model uncertainty for electronic health records. *ACM CHIL 2020 - Proceedings of the 2020 ACM Conference on Health, Inference, and Learning*, (pp. 204–213).

[50] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (pp. 214–226). New York, NY, USA: Association for Computing Machinery.

[51] D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., Hormozdiari, F., Houlsby, N., Hou, S., Jerfel, G., Karthikesalingam, A., Lucic, M., Ma, Y., McLean, C., Mincu, D., Mitani, A., Montanari, A., Nado, Z., Natarajan, V., Nielson, C., Osborne, T. F., Raman, R., Ramasamy, K., Sayres, R., Schrouff, J., Seneviratne, M., Sequeira, S., Suresh, H., Veitch, V., Vladymyrov, M., Wang, X., Webster, K., Yadlowsky, S., Yun, T., Zhai, X., & Sculley, D. (2020). Underspecification presents challenges for credibility in modern machine learning. *arXiv*.

[52] Elter, M., Schulz-Wendtland, R., & Wittenberg, T. (2007). The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process. *Medical Physics*, 34(11), 4164–4172.

[53] Farquhar, S., Osborne, M. A., & Gal, Y. (2020). Radial bayesian neural networks: Beyond discrete support in large-scale bayesian deep learning. In S. Chiappa & R. Calandra (Eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research* (pp. 1352–1362).: PMLR.

[54] Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, 27(8), 861–874.

[55] Fazelpour, S. & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass*, 16(8), e12760.

[56] Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(Vi).

[57] Franc, V. & Sonnenburg, S. (2008). Optimized cutting plane algorithm for support vector machines. In *Proceedings of the 25th International Conference on Machine Learning* (pp. 320–327).: ACM.

[58] Gal, Y. & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning.

[59] Gandy, O. H. (2010). Engaging rational discrimination: exploring reasons for placing regulatory constraints on decision support systems. *Ethics and Information Technology*, 12, 29–42.

[60] Ganesh, P., Chang, H., Strobel, M., & Shokri, R. (2023). On the impact of machine learning randomness on group fairness. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23 (pp. 1789–1800). New York, NY, USA: Association for Computing Machinery.

[61] Gass, S. & Saaty, T. (1955). The computational algorithm for the parametric objective function. *Naval research logistics quarterly*, 2(1-2), 39–45.

[62] Gentleman, R. & Lang, D. T. (2007). Statistical analyses and reproducible research. *Journal of Computational and Graphical Statistics*, 16(1), 1–23.

[63] Gepperth, A. & Hammer, B. (2016). Incremental learning algorithms and applications. In *European Symposium on Artificial Neural Networks (ESANN)* Bruges, Belgium.

[64] Giordano, R., Stephenson, W., Liu, R., Jordan, M., & Broderick, T. (2019). A swiss army infinitesimal jackknife. In K. Chaudhuri & M. Sugiyama (Eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research* (pp. 1139–1147).: PMLR.

[65] Gladwell, M. (2011). The order of things. *The New Yorker*.

[66] Goh, G., Cotter, A., Gupta, M., & Friedlander, M. P. (2016). Satisfying real-world goals with dataset constraints. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, volume 29: Curran Associates, Inc.

[67] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research* (pp. 1321–1330).: PMLR.

[68] Gurobi Optimization, LLC (2023). Gurobi Optimizer Reference Manual.

[69] Hamid, K., Asif, A., Abbasi, W., Sabih, D., et al. (2017). Machine learning with abstention for automated liver disease diagnosis. In *2017 International Conference on Frontiers of Information Technology (FIT)* (pp. 356–361).: IEEE.

[70] Hand, D. J. (1994). Deconstructing statistical questions. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 157(3), 317–338.

[71] Hand, D. J. (2006). Classifier Technology and the Illusion of Progress. *Statistical Science*, 21(1), 1 – 14.

[72] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*: NeurIPS.

[73] Hassein, N. (2017). Against black inclusion in facial recognition.

[74] Hofman, J. M., Goldstein, D. G., & Hullman, J. (2020). How Visualizing Inferential Uncertainty Can Mislead Readers about Treatment Effects in Scientific Results. *Conference on Human Factors in Computing Systems - Proceedings*.

[75] Hooker, S., Moorosi, N., Clark, G., Bengio, S., & Denton, E. (2020). Characterising bias in compressed models.

[76] Hsu, H. & Calmon, F. d. P. (2022). Rashomon capacity: A metric for predictive multiplicity in classification.

[77] Jacobs, A. Z. & Wallach, H. (2021). Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21 (pp. 375–385). New York, NY, USA: Association for Computing Machinery.

[78] Jiang, H. & Nachum, O. (2020). Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics* (pp. 702–712).: PMLR.

[79] Jiang, H., Narasimhan, H., Bahri, D., Cotter, A., & Rostamizadeh, A. (2022). Churn reduction via distillation. In *International Conference on Learning Representations*.

[80] Joachims, T., Finley, T., & Yu, C.-N. J. (2009). Cutting-plane training of structural SVMs. *Machine Learning*, 77(1), 27–59.

[81] Joseph, M., Kearns, M., Morgenstern, J. H., & Roth, A. (2016). Fairness in learning: Classic and contextual bandits. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, volume 29: Curran Associates, Inc.

[82] Joslyn, S. & LeClerc, J. (2013). Decisions With Uncertainty: The Glass Half Full. *Current Directions in Psychological Science*, 22(4), 308–315.

[83] Kale, A., Kay, M., & Hullman, J. (2020). Visual Reasoning Strategies for Effect Size Judgments and Decisions. *IEEE Transactions on Visualization and Computer Graphics*, (pp. 1–1).

[84] Keyes, O., Hutson, J., & Durbin, M. (2019). A mulching proposal: Analysing and improving an algorithmic system for turning the elderly into high-nutrient slurry. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems* (pp. 1–11).

[85] Khand, A., Frost, F., Grainger, R., Fisher, M., Chew, P., Mullen, L., Patel, B., Obeidat, M., Albouaini, K., Dodd, J., Goldstein, S. A., Newby, L. K., Cyr, D. D., Neely, M., Lüscher, T. F., Brown, E. B., White, H. D., Ohman, E. M., Roe, M. T., Hamm, C. W., Six, A. J., Backus, B. E., & Kelder, J. C. (2017). Heart Score Value. *Netherlands Heart Journal*, 10(6), 1–10.

[86] Kim, N. W., Bylinskii, Z., Borkin, M. A., Oliva, A., Gajos, K. Z., & Pfister, H. (2015). A crowdsourced alternative to eye-tracking for visualization understanding. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '15 (pp. 1349–1354). New York, NY, USA: Association for Computing Machinery.

[87] Kim, P. T. (2022). Race-aware algorithms: Fairness, nondiscrimination and affirmative action. *California law review*, 110, 1539.

[88] Kithulgoda, C. I., Vaithianathan, R., & Culhane, D. P. (2022). Predictive risk modeling to identify homeless clients at risk for prioritizing services using routinely collected data. *Journal of Technology in Human Services*, 40(2), 134–156.

[89] Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, 105(5), 491–495.

[90] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2022). The challenge of understanding what users want: Inconsistent preferences and engagement optimization. *arXiv preprint arXiv:2202.11776*.

[91] Kleinberg, J. M., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. In *Information Technology Convergence and Services*.

[92] Kohavi, R. (1996a). Census Income. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5GP7S.

[93] Kohavi, R. (1996b). Scaling up the accuracy of NB classifier : a DT hybrid. *Kdd*, (Utgoff 1988), 202–207.

[94] Kompa, B., Snoek, J., & Beam, A. L. (2021). Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1), 1–6.

[95] Kovačević, J. (2007). How to encourage and publish reproducible research. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '07*, ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings (pp. IV1273–IV1276). Copyright: Copyright 2011 Elsevier B.V., All rights reserved.; 2007 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '07 ; Conference date: 15-04-2007 Through 20-04-2007.

[96] Kristiadi, A., Hein, M., & Hennig, P. (2020). Being bayesian, even just a bit, fixes overconfidence in relu networks. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20: JMLR.org.

[97] Kulynych, B., Hsu, H., Troncoso, C., & Calmon, F. P. (2023). Arbitrary decisions are a hidden cost of differentially private training. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23 (pp. 1609–1623). New York, NY, USA: Association for Computing Machinery.

[98] Kurosawa, A. (1950). Rashomon. Motion Picture. Directed by Akira Kurosawa. Toho Co., Ltd.

[99] Lahoti, P., Gummadi, K. P., & Weikum, G. (2019). ifair: Learning individually fair data representations for algorithmic decision making. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)* (pp. 1334–1345).

[100] Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17 (pp. 6405–6416). Red Hook, NY, USA: Curran Associates Inc.

[101] Lan, X., Zhu, X., & Gong, S. (2018). Knowledge distillation by on-the-fly native ensemble. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18 (pp. 7528–7538). Red Hook, NY, USA: Curran Associates Inc.

[102] Latessa, E. J., Lemke, R., Makarios, M., Smith, P., & Lowenkamp, C. T. (2010). The creation and validation of the ohio risk assessment system (ORAS). *Federal Probation*, 74(1), 16–22.

[103] Liu, J., Zhong, C., Li, B., Seltzer, M., & Rudin, C. (2022). Fasterrisk: Fast and accurate interpretable risk scores. In *Neural Information Processing Systems (NeurIPS)*.

[104] Liu, J. Z., Lin, Z., Padhy, S., Tran, D., Bedrax-Weiss, T., & Lakshminarayanan, B. (2020). Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20 Red Hook, NY, USA: Curran Associates Inc.

[105] Long, C. X., Hsu, H., Alghamdi, W., & Calmon, F. (2023). Individual arbitrariness and group fairness. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[106] Lum, K., Dunson, D. B., & Johndrow, J. (2021). Closer than they appear: A Bayesian perspective on individual-level heterogeneity in risk assessment.

[107] Lum, K. & Isaac, W. (2016). To predict and serve? *Significance*, 13(5), 14–19.

[108] Mackay, D. J. C. (1992). *Bayesian Methods for Adaptive Models*. PhD thesis, California Institute of Technology, USA. UMI Order No. GAX92-32200.

[109] Macêdo, D. & Ludermir, T. (2022). Enhanced isotropy maximization loss: Seamless and high-performance out-of-distribution detection simply replacing the softmax loss.

[110] Malinin, A. & Gales, M. (2018). Predictive uncertainty estimation via prior networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18 (pp. 7047–7058). Red Hook, NY, USA: Curran Associates Inc.

[111] Martin Jr, D., Prabhakaran, V., Kuhlberg, J., Smart, A., & Isaac, W. S. (2020). Participatory problem formulation for fairer machine learning through community based system dynamics. *arXiv preprint arXiv:2005.07572*.

[112] Marx, C., Calmon, F. P., & Ustun, B. (2019). Predictive multiplicity in classification.

[113] Masana, M., Liu, X., Twardowski, B., Menta, M., Bagdanov, A. D., & van de Weijer, J. (2023). Class-incremental learning: Survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(05), 5513–5533.

[114] McGrath, S., Mehta, P., Zytek, A., Lage, I., & Lakkaraju, H. (2020). When Does Uncertainty Matter?: Understanding the Impact of Predictive Uncertainty in ML Assisted Decision Making.

[115] McNutt, M. (2014). Reproducibility. *Science*, 343(6168), 229–229.

[116] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2018). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1808.00023*.

[117] Mei, S. & Montanari, A. (2022). The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4), 667–766.

[118] Mesirov, J. P. (2010). Accessible reproducible research. *Science*, 327(5964), 415–416.

[119] Meyer, A. P., Albarghouthi, A., & D'Antoni, L. (2023). The dataset multiplicity problem: How unreliable data impacts predictions. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23 (pp. 193–204). New York, NY, USA: Association for Computing Machinery.

[120] Milli, S., Belli, L., & Hardt, M. (2021). From optimizing engagement to measuring value. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 714–722).

[121] Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8(1), 141–163.

[122] Moreno, R. P., Metnitz, P. G., Almeida, E., Jordan, B., Bauer, P., Campos, R. A., Iapichino, G., Edbrooke, D., Capuzzo, M., & Le Gall, J. R. (2005). SAPS 3 - From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. *Intensive Care Medicine*, 31(10), 1345–1355.

[123] Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62, 22–31.

[124] Mozannar, H. & Sontag, D. (2020). Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning* (pp. 7076–7087).: PMLR.

[125] Mullainathan, S. & Obermeyer, Z. (2021). On the inequity of predicting a while hoping for b. In *AEA Papers and Proceedings*, volume 111 (pp. 37–42).

[126] Naeini, M. P., Cooper, G. F., & Hauskrecht, M. (2015). Binary classifier calibration using a Bayesian non-parametric approach. *SIAM International Conference on Data Mining 2015, SDM 2015*, (pp. 208–216).

[127] Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., & Sutskever, I. (2019). Deep double descent: Where bigger models and more data hurt.

[128] Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. Berlin, Heidelberg: Springer-Verlag.

[129] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453.

[130] Oh, S. & Kumar, S. (2022). Robustness of deep recommendation systems to untargeted interaction perturbations. *CoRR*, abs/2201.12686.

[131] Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., & Snoek, J. (2019). *Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift*. Red Hook, NY, USA: Curran Associates Inc.

[132] Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural Networks*, 113, 54–71.

[133] Passi, S. & Barocas, S. (2019). Problem formulation and fairness. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, (pp. 39–48).

[134] Pawelczyk, M., Broelemann, K., & Kasneci, G. (2020). On counterfactual explanations under predictive multiplicity. *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence, UAI 2020*, 124, 839–848.

[135] Pedreshi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08 (pp. 560–568). New York, NY, USA: Association for Computing Machinery.

[136] Peng, R. D. (2011). Reproducible research in computational science. *Science*, 334(6060), 1226–1227.

[137] Pfohl, S., Xu, Y., Foryciarz, A., Ignatiadis, N., Genkins, J., & Shah, N. (2022). Net benefit, calibration, threshold selection, and training objectives for algorithmic fairness in healthcare. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1039–1052).

[138] Polikar, R., Upda, L., Upda, S. S., & Honavar, V. (2001). Learn++: An incremental learning algorithm for supervised neural networks. *Trans. Sys. Man Cyber Part C*, 31(4), 497–508.

[139] Provost, F. & Fawcett, T. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media, Inc.

[140] Qian, S., Pham, V. H., Lutellier, T., Hu, Z., Kim, J., Tan, L., Yu, Y., Chen, J., & Shah, S. (2021). Are my deep learning systems fair? an empirical study of fixed-seed training. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in Neural Information Processing Systems*, volume 34 (pp. 30211–30227).: Curran Associates, Inc.

[141] Rader, E., Cotter, K., & Cho, J. (2018). Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18 (pp. 1–13). New York, NY, USA: Association for Computing Machinery.

[142] Riquelme, C., Tucker, G., & Snoek, J. (2018). Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. In *International Conference on Learning Representations*.

[143] Romano, Y., Barber, R. F., Sabatti, C., & Candès, E. (2020). With Malice Toward None: Assessing Uncertainty via Equalized Coverage. *Harvard Data Science Review*, (pp. 1–14).

[144] Roth, A., Tolbert, A., & Weinstein, S. (2022). Reconciling individual probability forecasts.

[145] Rule, A., Birmingham, A., Zuniga, C., Altintas, I., Huang, S.-C., Knight, R., Moshiri, N., Nguyen, M. H., Rosenthal, S. B., Pérez, F., & Rose, P. W. (2018). Ten simple rules for reproducible research in jupyter notebooks.

[146] Semenova, L., Chen, H., Parr, R., & Rudin, C. (2023). A path to simpler models starts with noise. In *Proceedings of Neural Information Processing Systems (NeurIPS)*.

[147] Semenova, L., Rudin, C., & Parr, R. (2019). A study in Rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. (pp. 1–64).

[148] Semenova, L., Rudin, C., & Parr, R. (2022). On the existence of simpler machine learning models. In *ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*.

[149] Sensoy, M., Kaplan, L., & Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18 (pp. 3183–3193). Red Hook, NY, USA: Curran Associates Inc.

[150] Shafer, G. & Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9, 371–421.

[151] Shen, Y., Xiong, Y., Xia, W., & Soatto, S. (2020). Towards backward-compatible representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6367–6376). Los Alamitos, CA, USA: IEEE Computer Society.

[152] Shu, L., Xu, H., & Liu, B. (2017). DOC: Deep open classification of text documents. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2911–2916). Copenhagen, Denmark: Association for Computational Linguistics.

[153] Simonite, T. (2019). Algorithms allegedly penalized black renters. The US government is watching. Wired. Accessed: 2024-04-02.

[154] Snoek, J., Rippel, O., Swersky, K., Kiros, R., Satish, N., Sundaram, N., Patwary, M. M. A., Prabhat, P., & Adams, R. P. (2015). Scalable bayesian optimization using deep neural networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15 (pp. 2171–2180).: JMLR.org.

[155] Song, G. & Chai, W. (2018). Collaborative learning for deep neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18 (pp. 1837–1846). Red Hook, NY, USA: Curran Associates Inc.

[156] Sonnenburg, S., Braun, M. L., Ong, C. S., Bengio, S., Bottou, L., Holmes, G., LeCun, Y., Müller, K.-R., Pereira, F., Rasmussen, C. E., Rätsch, G., Schölkopf, B., Smola, A., Vincent, P., Weston, J., & Williamson, R. (2007). The need for open source software in machine learning. *J. Mach. Learn. Res.*, 8, 2443–2466.

[157] Soyer, E. & Hogarth, R. M. (2012). The illusion of predictability: How regression statistics mislead experts. *International Journal of Forecasting*, 28(3), 695–711.

[158] Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackermann, Z., et al. (2020). A deep learning approach to antibiotic discovery. *Cell*, 180(4), 688–702.

[159] Tagasovska, N. & Lopez-Paz, D. (2019). *Single-Model Uncertainties for Deep Learning*. Red Hook, NY, USA: Curran Associates Inc.

[160] Than, M., Flaws, D., Sanders, S., Doust, J., Glasziou, P., Kline, J., Aldous, S., Troughton, R., Reid, C., Parsonage, W. A., Frampton, C., Greenslade, J. H., Deely, J. M., Hess, E., Sadiq, A. B., Singleton, R., Shopland, R., Vercoe, L., Woolhouse-Williams, M., Ardagh, M., Bossuyt, P., Bannister, L., & Cullen, L. (2014). Development and validation of the emergency department assessment of chest pain score and 2h accelerated diagnostic protocol. *EMA - Emergency Medicine Australasia*, 26(1), 34–44.

[161] Tran, D., Liu, J., Dusenberry, M. W., Phan, D., Collier, M. P., Ren, J. J., Han, K., Wang, Z., Mariet, Z., Hu, C. H., Band, N., Rudner, T. G. J., Singhal, K., Nado, Z., van Amersfoort, J., Kirsch, A. C., Jenatton, R., Thain, N., Yuan, H., Buchanan, K., Murphy, K. P., Sculley, D., Gal, Y., Ghahramani, Z., Snoek, J. R., & Lakshminarayanan, B. (2022). Plex: Towards reliability using pretrained large model extensions. In *ICML Workshop: Principles of Distribution Shift (PODS)*.

[162] Träuble, F., Kügelgen, J. V., Kleindessner, M., Locatello, F., Schölkopf, B., & Gehler, P. V. (2021). Backward-compatible prediction updates: A probabilistic approach. In A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in Neural Information Processing Systems*.

[163] Ustun, B. & Rudin, C. (2017). Optimized Risk Scores. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: ACM.

[164] Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, (pp. 10–19).

[165] Ustun, B., Westover, M. B., Rudin, C., & Bianchi, M. T. (2016). Clinical prediction models for sleep apnea: The importance of medical history over symptoms. *Journal of Clinical Sleep Medicine*, 12(2), 161–168.

[166] van Amersfoort, J., Smith, L., Teh, Y. W., & Gal, Y. (2020). Uncertainty estimation using a single deep deterministic neural network.

[167] Vanschoren, J., Braun, M. L., & Ong, C. S. (2014). Open science in machine learning.

[168] Veitch, V., D'Amour, A., Yadlowsky, S., & Eisenstein, J. (2021). Counterfactual invariance to spurious correlations: Why and how to pass stress tests. *arXiv preprint arXiv:2106.00545*.

[169] Ventures, A. (2022). What is the psa?

[170] Viégas, Fernanda, B. & Wattenberg, M. (2006). Communication-minded visualization: A call to action. *IBM Systems Journal*, 45(4), 801–812. Copyright - Copyright International Business Machines Corporation Oct-Dec 2006; Document feature - Illustrations; ; Last updated - 2023-11-24; CODEN - IBMSA7; SubjectsTermNotLitGenreText - United States–US.

[171] Wang, Y., Wang, X., Beutel, A., Prost, F., Chen, J., & Chi, E. H. (2021a). Understanding and improving fairness-accuracy trade-offs in multi-task learning. *CoRR*, abs/2106.02705.

[172] Wang, Z. J., Kale, A., Nori, H., Stella, P., Nunnally, M., Chau, D. H., Vorvoreanu, M., Vaughan, J. W., & Caruana, R. (2021b). Gam changer: Editing generalized additive models with interactive visualization. In *Advances in Neural Information Processing Systems, Bridging the Gap: From Machine Learning Research to Clinical Practice (Research2Clinics) Workshop*.

[173] Wang, Z. J., Zhong, C., Xin, R., Takagi, T., Chen, Z., Chau, D. H., Rudin, C., & Seltzer, M. (2022). Timbertrek: Exploring and curating sparse decision trees with interactive visualization. In *2022 IEEE Visualization and Visual Analytics (VIS)* (pp. 60–64).: IEEE.

[174] Watson-Daniels, J., Barocas, S., Hofman, J. M., & Chouldechova, A. (2023a). Multi-target multiplicity: Flexibility and fairness in target specification under resource constraints. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23 (pp. 297–311). New York, NY, USA: Association for Computing Machinery.

[175] Watson-Daniels, J., du Pin Calmon, F., D'Amour, A., Long, C., Parkes, D. C., & Ustun, B. (2024). Predictive churn with the set of good models. arXiv:2402.07745.

[176] Watson-Daniels, J., Parkes, D. C., & Ustun, B. (2023b). Predictive multiplicity in probabilistic classification. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23: AAAI Press.

[177] Wei, D., Nair, R., Dhurandhar, A., Varshney, K. R., Daly, E. M., & Singh, M. (2022). On the safety of interpretable machine learning: A maximum deviation approach.

[178] Wieringa, M. (2020). What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20 (pp. 1–18). New York, NY, USA: Association for Computing Machinery.

[179] Xie, Y., an Lai, Y., Xiong, Y., Zhang, Y., & Soatto, S. (2021). Regression bugs are in your model! measuring, reducing and analyzing regressions in nlp model updates.

[180] Xin, R., Zhong, C., Chen, Z., Takagi, T., Seltzer, M., & Rudin, C. (2022). Exploring the whole rashomon set of sparse decision trees. In *Advances in Neural Information Processing Systems*, volume 35 (pp. 14071–14084).

[181] Yan, S., Xiong, Y., Kundu, K., Yang, S., Deng, S., Wang, M., Xia, W., & Soatto, S. (2021). Positive-congruent training: Towards regression-free model updates. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 14294–14303). Los Alamitos, CA, USA: IEEE Computer Society.

[182] Yeh, I. C. & Lien, C. h. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2 PART 1), 2473–2480.

[183] Zeng, J., Ustun, B., & Rudin, C. (2017). Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(3), 689–722.

[184] Zhang, Y., Xiang, T., Hospedales, T. M., & Lu, H. (2017). Deep mutual learning.

[185] Zhong, C., Chen, Z., Liu, J., Seltzer, M., & Rudin, C. (2023). Exploring and interacting with the set of good sparse generalized additive models. In *Thirty-seventh Conference on Neural Information Processing Systems*.