DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY

Analytics for an Online Retailer: Demand Forecasting and Price Optimization

Citation
Ferreira, Kris J., Bin Hong Alex Lee, and David Simchi-Levi. "Analytics for an Online Retailer: Demand Forecasting and Price Optimization." Manufacturing & Service Operations Management (forthcoming).

Published version
https://doi.org/10.1287/msom.2015.0561

Link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:26964422

Terms of use
This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at
https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFlOWI2M2EwYTg

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

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

Analytics for an Online Retailer: Demand Forecasting and Price Optimization

Kris Johnson Ferreira
Technology and Operations Management Unit, Harvard Business School, kferreira@hbs.edu

Bin Hong Alex Lee
Engineering Systems Division, Massachusetts Institute of Technology, binhong@mit.edu

David Simchi-Levi
Engineering Systems Division, Department of Civil & Environmental Engineering and the Operations Research Center, Massachusetts Institute of Technology, dslevi@mit.edu

We present our work with an online retailer, Rue La La, as an example of how a retailer can use its wealth of data to optimize pricing decisions on a daily basis. Rue La La is in the online fashion sample sales industry, where they offer extremely limited-time discounts on designer apparel and accessories. One of the retailer's main challenges is pricing and predicting demand for products that it has never sold before, which account for the majority of sales and revenue. To tackle this challenge, we use machine learning techniques to estimate historical lost sales and predict future demand of new products. The nonparametric structure of our demand prediction model, along with the dependence of a product's demand on the price of competing products, pose new challenges on translating the demand forecasts into a pricing policy. We develop an algorithm to efficiently solve the subsequent multi-product price optimization that incorporates reference price effects, and we create and implement this algorithm into a pricing decision support tool for Rue La La's daily use. We conduct a field experiment and find that sales does not decrease due to implementing tool recommended price increases for medium and high price point products. Finally, we estimate an increase in revenue of the test group by approximately 9.7% with an associated 90% confidence interval of [2.3%, 17.8%].

1. Introduction

We present our work with an online retailer, Rue La La, as an example of how a retailer can use its wealth of data to optimize pricing decisions on a daily basis. Rue La La is in the online fashion sample sales industry, where they offer extremely limited-time discounts ("flash sales") on designer apparel and accessories. According to McKitterick (2015), this industry emerged in the mid-2000s and by 2015 was worth approximately 3.8 billion USD, benefiting from an annual industry growth of approximately 17% over the last 5 years. Rue La La has approximately 14% market share in this industry, which is third largest to Zulily (39%) and Gilt Groupe (18%). Several of its smaller competitors also have brick-and-mortar stores, whereas others like Rue La La only sell products online. For an overview of the online fashion sample sales and broader "daily deal" industries, see Wolverson (2012), LON (2011), and Ostapenko (2013).

Upon visiting Rue La La's website (www.ruelala.com), the customer sees several "events", each representing a collection of for-sale products that are similar in some way. For example, one event

Figure 1 **Example of three events shown on Rue La La's website**

Figure 2 **Example of three styles shown in the men's sweater event**

might represent a collection of products from the same designer, whereas another event might represent a collection of men's sweaters. Figure 1 shows a snapshot of three events that have appeared on their website. At the bottom of each event, there is a countdown timer informing the customer of the time remaining until the event is no longer available; events typically last between 1-4 days.

When a customer sees an event he is interested in, he can click on the event which takes him to a new page that shows all of the products for sale in that event; each product on this page is referred to as a "style". For example, Figure 2 shows three styles available in a men's sweater event (the first event shown in Figure 1). Finally, if the customer likes a particular style, he may click on the style which takes him to a new page that displays detailed information about the style, including which sizes are available; we will refer to a size-specific product as an "item" or "SKU". The price for each item is set at the style level, where a style is essentially an aggregation of all sizes of otherwise identical items. Currently, the price does not change throughout the duration of the event.

Figure 3 highlights a few aspects of Rue La La's procure-to-pay process that are critical in understanding the work presented in this paper. First, Rue La La's merchants procure items from designers who typically ship the items immediately to Rue La La's warehouse[1]. On a frequent periodic basis, merchants identify opportunities for future events based on available styles in inventory, customer needs, etc. When the event starts, customers place orders, and Rue La La ships items from its warehouse to the customers. When the event ends or an item runs out of inventory, customers may no longer place an order for that item. If there is remaining inventory at the end of the event, then the merchants will plan a subsequent event where they will sell the same style[2]. We will refer to styles being sold for the first time as "first exposure styles"; a majority of Rue La

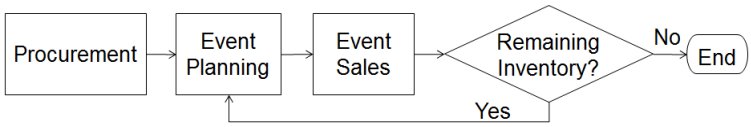

Figure 3 **Subset of Rue La La's procure-to-pay process**

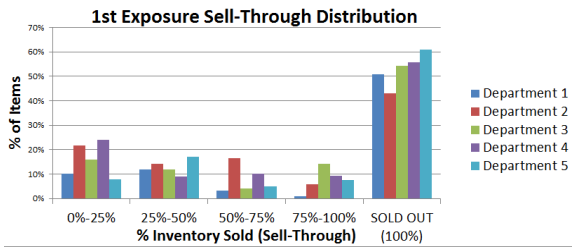

Figure 4 **First exposure (new product) sell-through distribution by department**

La's revenue comes from first exposure styles, and hundreds of first exposure styles are offered on a daily basis.

One of Rue La La's main challenges is pricing and predicting demand for these first exposure styles. Figure 4 shows a histogram of the sell-through (% of inventory sold) distribution for first exposure items in Rue La La's top 5 departments (with respect to quantity sold). For example, 51% of first exposure items in Department 1 sell out before the end of the event, and 10% sell less than 25% of their inventory. Department names are hidden and data disguised in order to protect confidentiality. Since a large percentage of first exposure items sell out before the sales period is over, it may be possible to raise prices on these items while still achieving high sell-through; on the other hand, many first exposure items sell less than half of their inventory by the end of the sales period, suggesting that the price may have been too high. These observations motivate the development of a pricing decision support tool, allowing Rue La La to take advantage of available data in order to maximize revenue from first exposure sales.

Our approach is two-fold and begins with developing a demand prediction model for first exposure items; we then use this demand prediction data as input into a price optimization model to maximize revenue. The two biggest challenges faced when building our demand prediction model are estimating lost sales due to stockouts, and predicting demand for items that have no historical sales data. We use machine learning techniques to address these challenges and predict future demand. Regression trees - an intuitive, yet nonparametric regression model - are shown to be effective predictors of demand in terms of both predictability and interpretability.

We then formulate a price optimization model to maximize revenue from first exposure styles, using demand predictions from the regression trees as inputs. In this case, the biggest challenge we face is that each style's demand depends on the price of competing styles, which restricts us from solving a price optimization problem individually for each style and leads to an exponential number

of variables. Furthermore, the nonparametric structure of regression trees makes this problem particularly difficult to solve. We develop a novel reformulation of the price optimization problem by exploiting a particular reference price metric, and we create and implement an efficient algorithm that allows Rue La La to optimize prices on a daily basis for the next day's sales. We conduct a field experiment and find that sell-through does not decrease due to implementing tool recommended price increases for medium and high price point styles. Furthermore, we estimate an increase in revenue of the test group by approximately 9.7% with an associated 90% confidence interval of [2.3%, 17.8%], significantly impacting their bottom line.

In the remainder of this section, we provide a literature review on related research and describe Rue La La's legacy pricing process. Section 2 includes details on the demand prediction model, while Section 3 describes the price optimization model and the efficient algorithm we developed to solve it. Details on the implementation of our pricing decision support tool as well as an analysis of the impact of our tool via field experiments are included in Section 4. Finally, Section 5 concludes the paper with a summary of our results and potential areas for future work.

1.1. Literature Review

There has been significant research conducted on price-based revenue management over the past few decades; see Özer and Phillips (2012) and Talluri and Van Ryzin (2005) for an excellent in-depth overview of such work. The distinguishing features of our work in this field include (*i*) the development and implementation of a pricing decision support tool for an online retailer offering "flash sales", including a field experiment that estimates the impact of the tool, (*ii*) the creation of a new model and efficient algorithm to set initial prices by solving a multi-product static price optimization that incorporates reference price effects, and (*iii*) the use of a nonparametric multi-product demand prediction model.

A group of researchers have worked on the development and implementation of pricing decision support tools for retailers. For example, Caro and Gallien (2012) implement a markdown multi-product pricing decision support tool for fast-fashion retailer, Zara; markdown pricing is common in fashion retailing where retailers aim to sell all of their inventory by the end of relatively short product life cycles. Smith and Achabal (1998) provide another example of the development and implementation of a markdown pricing decision support tool. Other pricing decision support tools focus on recommending promotion pricing strategies (e.g. see Natter et al. (2007) and Wu et al. (2014)); promotion pricing is common in consumer packaged goods to increase demand of a particular brand. Over the last decade, several software firms have introduced revenue management software to help retailers make pricing decisions; much of the available software currently focuses on promotion and markdown price optimization. Academic research on retail price-based revenue management also focuses on promotion and markdown dynamic price optimization. Özer

and Phillips (2012), Talluri and Van Ryzin (2005), Elmaghraby and Keskinocak (2003), and Bitran and Caldentey (2003) provide a good overview of this literature.

Rue La La's flash sales business model is not well-suited for dynamic price optimization and is thus unable to benefit from these advances in research and software tools. There are several characteristics of the online flash sales industry that make a single-price, static model more applicable. For example, many designers require that Rue La La limit the frequency of events which sell their brand in order not to degrade the value of the brand. Even without such constraints placed by designers, flash sales businesses usually do not show the same styles too frequently in order to increase scarcity and entice customers to visit their site on a daily basis, inducing myopic customer behavior. Therefore, any unsold items at the end of an event are typically held for some period of time before another event is created to sell the leftover items. To further complicate future event planning, purchasing decisions for new styles that would compete against today's leftover inventory have typically not yet been made. Ostapenko (2013) provides an overview of this industry's characteristics.

Since the popularity and competitive landscape for a particular style in the future - and thus future demand and revenue - is very difficult to predict, a single-price model that maximizes revenue given the current landscape is appropriate. Relatively little research has been devoted to multi-product single-price optimization models in the retail industry. Exceptions include work by Little and Shapiro (1980) and Reibstein and Gatignon (1984) that highlight the importance of concurrently pricing competing products in order to maximize the profitability of the entire product line. Birge et al. (1998) determine optimal single-price strategies of two substitutable products given capacity constraints, Maddah and Bish (2007) analyze both static pricing and inventory decisions for multiple competing products, and Choi (2007) addresses the issue of setting initial prices of fashion items using market information from pre-season sales.

In the operations management literature, aggregate demand is often modeled as a parametric function of price and possibly other marketing variables. See Talluri and Van Ryzin (2005) for an overview of multi-product demand functions that are typically used in retail price optimization. One reason for the popularity of these demand functions as an input to price optimization is their set of properties, such as linearity, concavity and increasing differences, that leads to simpler, tractable optimization problems that can provide managerial insights. In Section 2, we will be testing some of these functions as possible forecasting models using Rue La La's data. In addition, we chose not to initially restrict ourselves to the type of demand functions that would lead to simpler, tractable price optimization problems in hopes to achieve better demand predictions. We show that in fact a nonparametric demand prediction model works very well in this setting, and we resolve the structural challenges that this introduces to the price optimization problem.

There is also a large stream of work on individual consumer choice models which can be aggregated to estimate total demand. The most common of these are parametric, random utility models which model the utility that each consumer gains when making a purchase; see Talluri and Van Ryzin (2005) for an overview of these models in operations management and Berry et al. (1995) for the basis of a popular model in industrial organization. We chose to focus on aggregate demand models rather than random utility models because (*i*) the parameter estimation requirements for the random utility models - especially those incorporating substitution effects - are prohibitive in our situation, and (*ii*) each customer's choice set is constantly changing and difficult to define. To put these issues in context, in contrast to most retailers, Rue La La's assortment changes an average of twice a day when a new event begins, and the average inventory of each SKU in the assortment is less than 10 units. Although our aggregate demand model is not grounded in consumer utility theory, we believe that it is effective at predicting demand for Rue La La.

1.2. Legacy Pricing Process

For many retailers, initial prices are typically based on some combination of the following criteria: percentage markup on cost, competitors' pricing, and the merchants' judgement/feel for the best price of the product (see Subrahmanyan (2000), Levy et al. (2004), and Şen (2008)). These techniques are quite simple and none require style-level demand forecasting.

Rue La La typically applies a fixed percentage markup on cost for each of its styles and compares this to competitors' pricing of identical styles on the day of the event; they choose whichever price is lowest (between fixed markup and competitors' prices) in order to guarantee that they offer the best deal in the market to their customers. Interestingly, Rue La La usually does not find identical products for sale on other competitors' websites, and thus the fixed percentage markup is applied. This is simply because other flash sales sites are unlikely to be selling the same product on the same day (if ever), and other retailers are unlikely to offer deep discounts of the same product at the same time. A recent customer survey that Rue La La conducted shows that only approximately 15% of their customers occasionally comparison shop on other websites.

In the rest of this paper, we show that applying machine learning and optimization techniques to these initial pricing decisions - while maintaining Rue La La's value proposition - can have a substantial financial impact on the company.

2. Demand Prediction Model

A key requirement for our pricing decision support tool is the ability to accurately predict demand. To do so, the first decision we need to make is the level of detail in which to aggregate our forecasts. We chose to aggregate items at the style level - essentially aggregating all sizes of an otherwise identical product - and predicted demand for each style. The main reason for doing so is because

the price is the same within a style regardless of the size. Also, further aggregating styles into a more general level of the product hierarchy would eliminate our ability to measure the effects of competing styles' prices and presence in the assortment on the demand of each style.

One challenge we must address that arises from this decision is how to apply inventory constraints, since inventory is held for each size of the style. Let $\mathcal{I}$ denote the set of all styles that Rue La La sells, and given style i, let $\mathcal{S}(i)$ denote the set of all sizes associated with style i with $s \in \mathcal{S}(i)$ being a specific size. A unique item is thus represented by a pair $(i, s) \in \mathcal{I} \times \mathcal{S}(i)$ which we abbreviate as is. The demand for a style given that there are no inventory constraints is denoted u_i; the demand for a particular size is u_{is}. In reality, Rue La La has very limited inventory of many of their products. Inventory is held in each size of a style and is denoted by C_{is}; total inventory for style i is $C_i = \sum_{s \in S(i)} C_{is}$. Sales for each size of a style is therefore constrained by inventory and is defined as $d_{is} = \min\{C_{is}, u_{is}\}$. Sales for the entire style is defined as $d_i = \sum_{s \in S(i)} d_{is}$.

The following steps outline our demand prediction approach:

1. Record sales, d_i, of styles sold in the past.
2. Estimate demand, u_i, of styles sold in the past.
3. Predict demand and sales of new styles to be sold in the future, denoted $\hat{u}_i$ and $\hat{d}_i$, respectively.

Section 2.1 outlines the available data including sales of styles sold in the past, d_i, and summarizes the features, i.e. explanatory variables, that we developed from this data. Section 2.2 describes how we estimate demand, u_i, of styles sold in the past. Section 2.3 explains the development of our regression model that predicts demand, $\hat{u}_i$, and sales, $\hat{d}_i$, of future styles. Section 2.4 concludes by offering insights as to why the selected demand prediction model works well in our setting, and highlights the implications of our model on the subsequent task of price optimization.

2.1. Data for Demand Prediction Model

We were provided with sales transactions data from the beginning of 2011 through mid-2013, where each data record represents a time-stamped sale of an item during a specific event. This data includes the quantity sold of each SKU (d_{is}), price, event start date/time, event length, and the initial inventory of the item. In addition, we were provided with product-related data such as the product's brand, size, color, MSRP (manufacturer's suggested retail price), and hierarchy classification. With regards to hierarchy classification, each item aggregates (across all sizes) to a *style*, styles aggregate to form *subclasses*, subclasses aggregate to form *classes*, and classes aggregate to form *departments*.

Determining potential predictors of demand that could be derived from this data was a collaborative process with our main contacts at Rue La La, the COO and the VP of Pricing & Operations Strategy. Together we developed the features for our demand prediction model summarized in Figure 5. We provide a description for each of the less intuitive features in Appendix A.

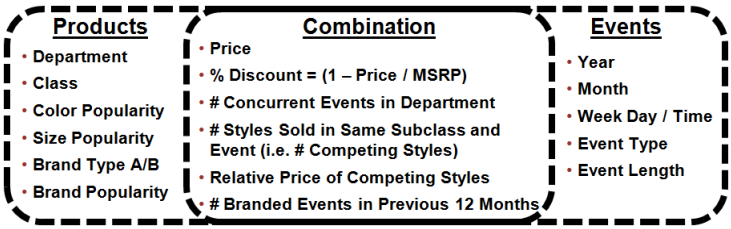

Figure 5 Summary of features used to develop demand prediction model

Three of the features are related to price. First, we naturally include the price of the style itself as a feature. Second, we include the percent discount off MSRP as one of our features. Finally, we include the relative price of competing styles, where we define competing styles as styles in the same subclass and event. This metric is calculated as the price of the style divided by the average price of all competing styles; this feature is meant to capture how a style's demand changes with the price of competing styles shown on the same page.

By including the relative price of competing styles as a feature, we are essentially considering the average price of competing styles as a *reference price* for consumers. Most of the research that has been conducted on the impact of reference prices on consumer decisions has been on internal reference prices for frequently purchased packaged goods; see Mazumdar et al. (2005) for a survey of the literature. The smaller body of research that has been conducted on the use of external reference prices for durable goods has found that current prices of competitive products and economic trends can be useful consumer reference prices (e.g. Winer (1985); Mazumdar et al. (2005)). We believe that the relative price of competing styles is an appropriate measure of a reference price that consumers can easily estimate in an online setting such as Rue La La's, where many competing products and associated prices are displayed on the same page. Emery (1970) suggests that such a metric is indeed used by consumers in price perception. Although we also would have liked to incorporate external reference prices, such as competitor's pricing, into our demand prediction model, we found that collecting and incorporating this data would be prohibitively difficult.

2.2. Estimating Demand for Styles Sold in the Past

When Rue La La sells out of a SKU prior to the end of the event, quantity sold, d_{is}, underestimates true demand, u_{is}, due to lost sales during the stockout period; in other words, $d_{is} \leq u_{is}$. Since we need to understand u_{is} to be able to predict demand of future styles, we must develop a way to estimate lost sales in historical data.

Almost all retailers face the issue of lost sales, and there has been considerable work done in this area to quantify the metric. See Section 9.4 in Talluri and Van Ryzin (2005) for an overview of common methods; some of these ideas are extended in Anupindi et al. (1998), Vulcano et al. (2012), and Musalem et al. (2010) to estimate lost sales for multiple, partially substitutable products. Each of these approaches requires a significant amount of sales data to estimate the parameters needed

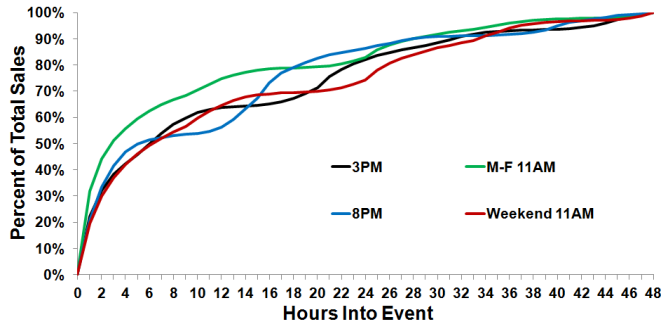

Figure 6 **Demand curves (percent of total sales by hour) for 2-day events**

to implement the method. While this data is available for many retailers, Rue La La operates in an extremely limited inventory environment (average SKU inventory is less than 10 units), making it infeasible to accurately estimate and apply these methods. Therefore, we chose to develop a different approach for Rue La La that utilizes product and event information as well as their knowledge of when each sale occurred during the event.

The idea of our method is to use sales data from items that did not sell out (i.e. when $u_{is} = d_{is}$) to estimate lost sales of items that did sell out. For each event length - ranging from 1 to 4 days - event start time of day, day of week, and department, we aggregated hourly sales over all items that did not sell out in the event. Then we calculated the percent of sales that occurred in each hour of the event which gave an empirical distribution of the proportion of sales that occur in the first X hours of an event, referred to internally as a "demand curve". This led to nearly 1,000 demand curves, although many of these curves were built from sales of just a few items. To aggregate and create the fewest distinct and interpretable demand curves, we performed hierarchical clustering on the proportion of sales that occurred in each hour. We briefly summarize the methodology that we used in Appendix B; see Everitt et al. (2011) for more details on cluster analysis.

Figure 6 shows the resulting categorizations that we made via clustering for 2-day events, as well as their associated demand curves. As the figure shows, each curve is relatively steep at the beginning of the event as well as the following day at 11:00am, when the majority of the events start. For example, an event that starts at 11:00am gets a second boost in sales at 11:00am the following day, 24 hours into the event; similarly, an event that starts at 8:00pm gets a second boost in sales at 11:00am the following day, this time only 15 hours into the event. Indeed, it appears that the driving force behind the shape of the demand curve is the customer traffic pattern on the site. The same procedure was followed for the other possible event lengths and similar insights were made. Interestingly, in Garro (2011)'s approach to estimating lost sales of Zara's products, they also find that the demand rate for each product is primarily a function of store traffic.

To estimate the demand for an item that did sell out, we identified the time that the item sold out and used the appropriate demand curve to estimate the proportion of sales that typically

occur within that amount of time. By simply dividing the number of units sold ($d_{is} = C_{is}$) by this proportion, we get an estimate of u_{is}. Demand for the style is estimated as $u_i = \sum_{s \in S(i)} u_{is}$. In Appendix B, we report on the accuracy of our method.

2.3. Predicting Demand and Sales for New Styles

We used the features and estimates of u_i from the historical data in order to build a regression model for each department that predicts demand of future first exposure styles, denoted $\hat{u}_i$. We built commonly used regression models, as well as others including regression trees, a nonparametric approach to predicting demand. We chose to focus on models which are interpretable for managers and merchants at Rue La La in order to better ensure the tool's adoption. Models tested included least squares regression, principal components regression, partial least squares regression, multiplicative (power) regression, semilogarithmic regression, and regression trees[3].

In order to compare multiple regression models, we first randomly split the data into training and testing data sets and used the training data to build the regression models. For those models requiring tuning parameters, we further used 5-fold cross-validation on the training data to set these parameters; please refer to Section 7.10 in Hastie et al. (2009) for details on cross-validation. The output of each regression model is a prediction of demand for each first exposure style, $\hat{u}_i$. Then we must transform $\hat{u}_i$ to $\hat{u}_{is}$ in order to apply inventory constraints. To do this, we used historical data and Rue La La's expertise to build size curves for each type of product; a size curve represents the percent of style demand that should be allocated to each size.

More formally, we use "product type" to denote similar products that have the same set of possible sizes, e.g. women's shoes vs. men's shirts. Let $\mathcal{T}$ be the set of all product types and $t(i) \in \mathcal{T}$ be the product type associated with style i. Let $\mathcal{S}(t(i))$ be the set of all possible sizes associated with product type $t(i)$, and observe that $\mathcal{S}(i) \subseteq \mathcal{S}(t(i))$. We let $s \in \mathcal{S}(t(i))$ be a specific size in $\mathcal{S}(t(i))$. A size curve for style i specifies the percent of $\hat{u}_i$ to allocate to size $s \in \mathcal{S}(t(i))$, denoted $q_{(t(i),s)}$. Note that $\sum_{s \in \mathcal{S}(t(i))} q_{(t(i),s)} = 1$, although $\sum_{s \in \mathcal{S}(i)} q_{(t(i),s)} \leq 1$. To estimate $\hat{u}_{is}$ given $\hat{u}_i$, we simply apply the following formula using the size curve: $\hat{u}_{is} = \hat{u}_i * q_{(t(i),s)} \ \forall \ s \in \mathcal{S}(i)$. Figure 7 summarizes how we transform the output of our demand prediction model, $\hat{u}_i$, to a prediction of sales, $\hat{d}_i$.

In Appendix C, we present our comparison of the different regression models for Rue La La's largest department and show how we evaluated model performance. Across all performance metrics evaluated, regression trees with bagging consistently outperformed the other regression models for all departments, so we briefly explain this technique and refer the reader to Hastie et al. (2009) for a detailed discussion. See Figure 8 for an illustrative example of a regression tree. To predict demand using this tree, begin at the top of the tree and ask "Is the price of this style less than \$100?". If yes, then move down the left side of the tree and ask "Is the relative price of competing styles less

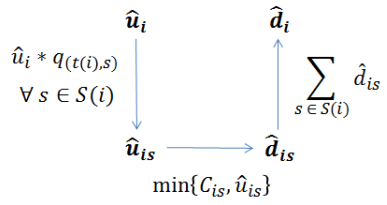

Figure 7 **Transforming demand prediction ($\hat{u}_i$) to sales prediction ($\hat{d}_i$)**

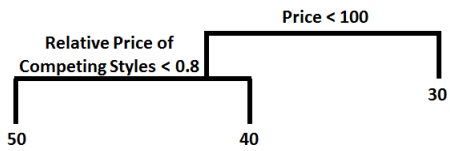

Figure 8 **Illustrative example of a regression tree**

than 0.8 (i.e. is the price of this style less than 80% of the average price of competing styles)?". If yes, then $\hat{u}_i = 50$; otherwise, $\hat{u}_i = 40$. Given this simple structure, regression trees are considered to be interpretable, especially to people with no prior knowledge of regression techniques.

One criticism of regression trees is that they are prone to overfitting from growing the tree too large; we addressed this issue in three ways. First, we required that a branch of the tree could only be split if the overall R^2 increased by the value of a "complexity parameter" found via cross-validation; for Rue La La's largest department, the complexity parameter was set at $\frac{1}{10000}$. Second, we limited the number of observations in a terminal node to be greater than a certain amount found via cross-validation. For Rue La La's largest department, the minimum number of observations in each terminal node is 10; the average number of observations is 21. To shed some more light on the structure of the 100 trees, the average number of binary splits was 581, and the average number of terminal nodes was 287.

Third, we employed a bootstrap aggregating ("bagging") technique to reduce the variance caused by a single regression tree. To do this, we randomly sampled N records from our training data (with replacement), where N is the number of observations in our training data; then we built a regression tree from this set of records. We did this 100 times in order to create 100 regression trees. To use these for prediction, we simply make 100 predictions of $\hat{u}_i$ using the 100 regression trees, transform each of these to sales predictions which are constrained by inventory, and take the average as our final sales prediction, $\hat{d}_i$.[4]

2.4. Discussion

We find it very interesting that regression trees with bagging outperformed other models we tested, including the most common models used in previous research in the field, on a variety of performance metrics. Talluri and Van Ryzin (2005) even comment that nonparametric forecasting

methods such as regression trees are typically not used in revenue management applications, primarily because they require a lot of data to implement and do a poor job extrapolating situations new to the business. In our case, Rue La La's business model permits a data-rich environment, and we place constraints on the set of possible prices for each style which helps avoid extrapolation issues; we describe these constraints as part of our implementation in Section 4.1. Thus, we believe that the concerns of such nonparametric models like regression trees are mitigated in this environment.

A benefit of using regression trees is that they do not require specification of a certain functional, parametric form between features and demand; the model is more general in this sense. In some respect, regression trees are able to determine - for each new style to be priced - the key characteristics of that style that will best predict demand, and they use *only* the demand of styles sold in the past that also had those same key characteristics as an estimate of future demand. In essence, regression trees are able to define and identify "similar" products sold in the past in order to help estimate future demand. Because of this, we think that regression trees could make effective demand prediction models for many new products - not just those sold in flash sales - and better help companies with pricing, production planning, etc. for new product introduction.

Another property of regression trees is that they do not require demand to be decreasing with price. Several empirical studies have been conducted that suggest demand may in fact increase with price for some products; for these products, price may be a signal of quality (see, e.g., Gaur and Fisher (2005)). The flexibility of regression trees provides a way to identify when price is a signal of quality and adjusts the demand forecast accordingly, sometimes even showing an increase in demand with rising prices. Along with fashion apparel, there are certainly types of products in other industries whose price could be a signal of quality; in these cases, regression trees or other nonparametric machine learning techniques may make effective demand prediction models since they allow for this non-monotonic relationship between demand and price.

While regression trees are effective in predicting demand, unfortunately their nonparametric structure leads to a more difficult price optimization problem. Another complication that arises in our price optimization is due to the "relative price of competing styles" feature. To justify the necessity of this feature, we evaluated its importance in the regression trees (commonly known as "variable importance"). The method we used to calculate variable importance of each feature was developed in Strobl et al. (2008) and has since been implemented in R. The method uses a permutation test that conditions on correlated features to test the conditional independence between the feature and response ($\hat{u}_i$). The percent increase in mean squared error due to the permutation gives the variable importance of the feature. Table 1 shows the features with the largest values of variable importance. The relative price of competing styles feature has the largest variable importance, and thus we believe it is necessary to include in our price optimization model.

Feature	Variable Importance
Relative price of competing styles	37.6
Discount	37.2
Number of competing styles	30.8
Brand MSRP Index	26.8
Color popularity	24.9
Price	22.2

Table 1 Features with largest variable importance

3. Price Optimization

Rue La La typically does not have the ability to buy inventory based on expected demand; rather, their purchasing decision is usually a binary "buy / no buy" decision based on what assortment of styles and sizes a designer may offer them. Because of this, our focus is on using the demand prediction model to determine an optimal pricing strategy that maximizes revenue, given pre-determined purchasing and assortment decisions.

As Figure 5 shows, several features are used to build the demand prediction models; however, only three of them are related to price - price, discount, and relative price of competing styles. Recall that the relative price of competing styles is calculated as the price of the style divided by the average price of all styles in the same subclass and event. One of the main challenges in converting the demand prediction models to a pricing decision support tool arises from this input, which implies that the demand of a given style depends on the price of competing styles. Thus, we cannot set prices of each style in isolation; instead, we need to make pricing decisions concurrently for all competing styles. Depending on the department, the number of competing styles could be as large as several hundred. An additional level of complexity stems from the fact that the relationship between demand and price is nonlinear and non-concave/convex due to the use of regression trees, making it difficult to exploit a well-behaved structure to simplify the problem.

Let $\mathcal{N}$ be the set of styles in a given subclass and event with $N = |\mathcal{N}|$ representing the number of styles, and let $\mathcal{M}$ be the set of possible prices for each style with $M = |\mathcal{M}|$ representing the number of possible prices. We assume without loss of generality that each style has the same set of possible prices. As is common for many discount retailers, Rue La La typically chooses prices that end in 4.90 or 9.90 (i.e. \$24.90 or \$119.90). The set of possible prices is characterized by a lower bound and an upper bound and every increment of 5 dollars inbetween; for example, if the lower bound of a style's price is \$24.90 and the upper bound is \$44.90, then the set of possible prices is $\mathcal{M} = \{\$24.90, \$29.90, \$34.90, \$39.90, \$44.90\}$, and $M = 5$. Let p_j represent the jth possible price in set $\mathcal{M}$, where $j = 1, ..., M$.

The objective of the price optimization problem is to select a price from $\mathcal{M}$ for each style in order to maximize the revenue earned from these styles in their first exposure. We recognize that

this is a myopic approach as opposed to maximizing revenue over the styles' lifetime. However, as described in Section 1.1, there are several characteristics of the online fashion sample sales industry that make this approach appropriate.

A naive approach to solving the problem would be to calculate and compare the expected revenue for each combination of possible prices assigned to each style. This requires predicting demand for each style given each competing style's price. Since there are M^N possible combinations of prices for all styles in a given subclass and event, this approach is often computationally intractable. Thus, we need a different approach that does not require considering all M^N price combinations. In the following section, we formulate our pricing problem as an integer program and develop an algorithm in Section 3.2 to efficiently price competing styles concurrently.

3.1. Integer Formulation

A key observation is that the relative price of competing styles feature dictates that the demand of a style is only dependent on the price of that style and the *sum* of the prices of all the competing styles, *not* the individual price of each style. Let k represent the sum of the prices of all competing styles. For example, consider a given subclass and event with $N = 3$ styles. All else equal, the relative price of competing styles for the first style would be the same for the price set {\$24.90, \$29.90, \$39.90} as it would be for the price set {\$24.90, \$34.90, \$34.90}; in each case, $k = \$94.70$, and the relative price of competing styles for the first style is $\frac{\$24.90}{k/3} = 0.79$. Let $\mathcal{K}$ denote the set of possible values of k, and note that the possible values of k range from $(N * \min_j\{p_j\})$ to $(N * \max_j\{p_j\})$ in increments of 5. We have $K \triangleq |\mathcal{K}| = N * (M-1) + 1$.

By considering the sum of prices of all competing styles rather than each style's price, we have effectively mapped the M^N possible price combinations to $O(MN)$ possible sums to consider. Using this set $\mathcal{K}$, one approach is to fix the value of k and solve an integer program to find the set of prices that maximizes revenue, requiring that the sum of prices chosen is k; then we could solve such an integer program for each possible value of $k \in \mathcal{K}$ and choose the solution with the maximum objective value. To do this, we define binary variables $x_{i,j}$ such that $x_{i,j} = 1$ if style i is assigned price p_j, and $x_{i,j} = 0$ otherwise, for all $i = 1, ..., N$ and $j = 1, ..., M$. The uncertainty in our model is given by $D_{i,j,k}$, a random variable representing sales of the ith style and j^{th} possible price when the sum of prices of competing styles is k.

We can write the following integer program, (IP_k):

$$\begin{aligned} \max & \sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{M}} p_j \mathbb{E}[D_{i,j,k} | p_j, k] x_{i,j} \\ st. \quad & \sum_{j \in \mathcal{M}} x_{i,j} = 1 \quad \forall\, i \in \mathcal{N} \\ & \sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{M}} p_j x_{i,j} = k \end{aligned}$$

$$x_{i,j} \in \{0,1\} \quad \forall\, i \in \mathcal{N},\ j \in \mathcal{M}$$

The first set of constraints guarantees that a single price is assigned to each style. The second constraint requires that the sum of prices over all styles must equal k. We can replace $\mathbb{E}[D_{i,j,k}|p_j,k]$ in the objective with its forecast from our demand prediction model described in Section 2, evaluated with p_j and k. Denote this forecast as $\tilde{D}_{i,j,k}$, which is identical to $\hat{d}_i$ but also clearly specifies feature information, p_j and k, since the demand forecast is made for different values of these features for the same style. Because our demand prediction model is built using regression trees, note that $D_{i,j,k}$ follows no particular functional form of p_j and k that exhibits properties such as linearity, concavity/convexity, etc. typically used in revenue management to simplify and solve problems.

In order to find the optimal set of price assignments that maximizes first exposure revenue, we solve (IP_k) for all $k \in \mathcal{K}$ and choose the solution with the maximum objective value, denoted as the solution to (IP) where $(IP) = \max_k(IP_k)$. When we first tested this approach at Rue La La, we found that the run-time was too long. Over a three month period, the average time it took to run all the (IP_k)s to price a day's worth of new products was nearly 6 hours with the longest time encountered being over three days; details of the hardware in which we ran the tool are given in Section 4.1. Since the tool must set prices on a daily basis, this length of time was unacceptable.

Next we present an efficient algorithm to solve this problem to optimality that exploits a special structure of our problem, namely that (IP_k) is very similar to the multiple-choice knapsack problem (MCKP) where each "class" in the MCKP corresponds to the price set $\mathcal{M}$. The only difference in formulations is that our constraint $\sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{M}} p_j x_{i,j} = k$ is an equality constraint instead of the inequality constraint found in the MCKP. The interpretation of the constraints in the MCKP, "exactly one item must be taken from each class, and the sum of the weights must not exceed k", can be translated to our setting as "exactly one price must be chosen for each style, and the sum of the prices must be exactly k".

3.2. Efficient Algorithm

Consider the following linear programming relaxation, (LP_k), to (IP_k):

$$\begin{aligned}
\max & \sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{M}} p_j \tilde{D}_{i,j,k} x_{i,j} \\
st. \quad & \sum_{j\in\mathcal{M}} x_{i,j} = 1 \quad \forall\, i \in \mathcal{N} \\
& \sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{M}} p_j x_{i,j} = k \\
& 0 \le x_{i,j} \le 1 \quad \forall\, i \in \mathcal{N},\ j \in \mathcal{M}
\end{aligned}$$

Let $z^*_{LP_k}$ and $z^*_{IP_k}$ be the optimal objective values for (LP_k) and (IP_k), respectively. The following theorem provides a bound on their difference.

THEOREM 1. *For any $k \in \mathcal{K}$,*

$$z^*_{LP_k} - z^*_{IP_k} \leq \max_{i \in \mathcal{N}} \left\{ \max_{j \in \mathcal{M}} \{p_j \tilde{D}_{i,j,k}\} - \min_{j \in \mathcal{M}} \{p_j \tilde{D}_{i,j,k}\} \right\} .$$

Note that for a given style, $\max_{j \in \mathcal{M}} \{p_j \tilde{D}_{i,j,k}\} - \min_{j \in \mathcal{M}} \{p_j \tilde{D}_{i,j,k}\}$ is simply the maximum difference in revenue across all possible prices. The outer maximization in Theorem 1 chooses this maximum difference across all styles. The key observation is that this bound is effectively constant and independent of problem size! To see this, consider a realistic maximum difference in revenue across all possible prices for any style in a given subclass and event, and use this as the bound.

The proof of Theorem 1 exploits the fact that $\mathcal{K}$ is constructed by taking summations over all possible prices of each style. This structure allows us to obtain such a result that the MCKP does not necessarily permit. We use this structure in an iterative method to show that there exists an optimal solution to (LP_k) that contains fractional variables relating to no more than one style; the technique has similarities to the iterative algorithm described in Lau et al. (2011). Furthermore, by showing the existence of a feasible integer solution by only changing this style's fractional variables to binary variables, we obtain the bound in the theorem. A detailed proof is provided in Appendix D.

Next we present our LP Bound Algorithm that utilizes Theorem 1 to solve (IP) to optimality. Let z^*_{IP} be the optimal objective value of (IP).

Algorithm 1 LP Bound Algorithm

1. For each possible value of k...
   (a) Solve (LP_k) to find $z^*_{LP_k}$.
   (b) Calculate the lower bound of $z^*_{IP_k}$, denoted as LB_k:

$$LB_k = z^*_{LP_k} - \max_{i \in \mathcal{N}} \left\{ \max_{j \in \mathcal{M}} \{p_j \tilde{D}_{i,j,k}\} - \min_{j \in \mathcal{M}} \{p_j \tilde{D}_{i,j,k}\} \right\}$$

2. Sort the possible values of $k \in \mathcal{K}$ in descending order according to $z^*_{LP_k}$. Let k_l represent the lth possible value of k in the ordered set; by construction, we have $z^*_{LP_{k_l}} \geq z^*_{LP_{k_{l+1}}}$.
3. Calculate a lower bound for the optimal objective value of (IP), denoted as LB: $LB = \max_{k \in \mathcal{K}} \{LB_k\}$. Set $\hat{k} = \arg\max_{k \in \mathcal{K}} \{LB_k\}$.
4. For $l = 1, ..., K$
   (a) Solve (IP_{k_l}).
   (b) If $z^*_{IP_{k_l}} > LB$, then set $\hat{k} = k_l$ and $LB = z^*_{IP_{\hat{k}}}$, and store the optimal solution.
   (c) If $LB \geq z^*_{LP_{k_{l+1}}}$ or $l = K$, then terminate the algorithm; otherwise, continue loop.

THEOREM 2. *The LP Bound Algorithm terminates with $z^*_{IP} = z^*_{IP_{\hat{k}}}$. An optimal solution to (IP) is the optimal set of price assignments given by the solution to $(IP_{\hat{k}})$.*

Number of (LP_k)s Solved	Number of (IP_k)s Solved	Optimality Gap After Solving All (LP_k)s
580	1	1.2%
569	1	1.4%
374	7	5.2%
321	2	1.9%
282	4	2.1%

Table 2 Example of LP Bound Algorithm Performance

The proof is straightforward and is presented in Appendix D. This algorithm can run in pseudo-polynomial time since it requires solving at most $K = N * (M - 1) + 1$ multiple-choice knapsack problems, each of which can be solved in pseudo-polynomial time (see Du and Pardalos (1998)).

The real benefit of this algorithm is that rather than solving K integer programs - (IP_K) $\forall$ $k = 1, ..., K$ - it solves K linear programming relaxations, one for each of the integer programs. It then uses the bound found in Theorem 1 to limit the number of integer programs that it needs to solve to find the optimal integer solution. Since the bound is independent of problem size, it is especially tight for large problems where the run-time could be an issue when solving K integer programs.

It is important to note that our LP Bound Algorithm does not guarantee that not all (IP_k)s will have to be solved. However, our analysis of the algorithm's performance over a three month test period indicates that the LP Bound Algorithm very rarely requires more than just a few integer programs to be run before reaching an optimal solution for (IP). For a subset of one day's subclass and event combinations, Table 2 shows the number of (LP_k)s solved (i.e. the number of possible values of k), the number of (IP_k)s solved in the LP Bound Algorithm, and the optimality gap after solving only the (LP_k)s; each record in the table represents data for a single subclass and event combination. As shown in the table, even with several hundred possible values of k, very few integer programs needed to be solved when using the LP Bound Algorithm. In a three month test period, we have not seen more than seven (IP_k)s solved when running our LP Bound Algorithm, highlighting the power of such a tight bound in Theorem 1.

Over the same three month period that we used to test the run-time of running all of the (IP_k)s, we calculated the average time it took to run the LP Bound Algorithm on a day's worth of new products as less than one hour with the longest time encountered being approximately 4.5 hours. Recall that this is a decrease from an average of nearly 6 hours and longest time of over three days when we ran only the (IP_k)s. If we wanted to guarantee that our algorithm would not take too long to solve, we could implement a stopping criteria that allows the algorithm to solve no more than X (IP_k)s; if this limit is reached, the algorithm could simply output the best integer solution calculated thus far.

4. Implementation and Results

Sections 2 and 3 described the demand prediction and price optimization models we developed to price Rue La La's first exposure styles. The first part of this section describes how we transformed these two models into a pricing decision support tool that we recently implemented at Rue La La. In Section 4.2, we present the design and results of a field experiment which show an expected increase in first exposure styles' revenue in the test group of approximately 9.7% with an associated 90% confidence interval of [2.3%, 17.8%], while minimally impacting aggregate sales.

4.1. Implementation

In the implementation of our pricing decision support tool, there are a few factors which we have to consider. The first is determining the price range for each style, i.e. the lower bound and upper bound of the set of possible prices described in Section 3. The lower bound on the price is the legacy price described in Section 1.2, denoted p_L; this is the minimum price that ensures Rue La La can earn a necessary profit per item and guarantees that they have the lowest price in the market. To maintain its market position as the low-cost provider of fashion products, Rue La La provided us a minimum discount percentage off MSRP for each department, which is an upper bound on the style's price. To further ensure the customer is getting a great deal, we restrict the upper bound to be no more than the maximum of \$15 or 15% greater than the lower bound. Thus, the upper bound on the possible set of prices is calculated as

$$\min\{(1 - \textit{minimum discount off MSRP}) * MSRP, \max\{\textit{lower bound} + 15, \textit{lower bound} * 1.15\}\}.$$

In the rare case that this upper bound is less than the lower bound, the upper bound is set equal to the lower bound, i.e. no price change is permitted.

Note that the implementation of our tool is such that it will only recommend price increases or no change in price. To maintain its low-cost position in the market, Rue La La only accepts the tool's recommended price increases if they are lower than what can be found on other websites.

A second factor to consider is the level of integration and automation in which to build the pricing decision support tool. The goal of the tool is to assist Rue La La in decision making without unduly burdening the executives or analysts with additional time and resources required for its execution. As such, after a period of close monitoring and validation of the tool's output, the pricing decision support tool has been implemented as a fully automated tool. It is run automatically every day, providing price recommendations to merchants for events starting the next day.

The entire pricing decision support tool is depicted in the architecture diagram in Figure 9. It consists of a Retail Price Optimizer ("RPO") which is our demand prediction and price optimization component. The input to RPO comes from Rue La La's Enterprise Resource Planning (ERP) system, which RPO is integrated with; the inputs consist of a set of features ("Impending Event

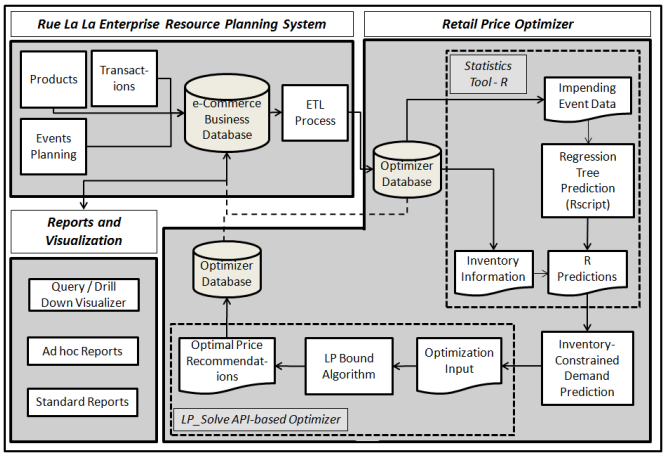

Figure 9 Architecture of pricing decision support tool

Data") which define the characteristics of the future event required to make demand predictions. This demand prediction process is performed using the statistical tool "R" (www.r-project.org); the 100 regression trees used to predict future demand for each department are stored in RPO, resulting in a more efficient way to make demand predictions. Inventory constraints are then applied to the predictions (from each regression tree) using inventory data for each item obtained from the database. The sales predictions are then fed to the LP Bound Algorithm to obtain an optimal pricing strategy for the event; this is implemented with the lp_solve API (lpsolve.sourceforge.net/5.5).

On average, the entire tool takes under one hour to run a day's worth of events; a typical run requires solving approximately 12 price optimization problems, one for each subclass and event combination for the next day's events. The longest run-time we encountered for a day's worth of events was 4.5 hours. These are reasonable run-times given Rue La La is running this tool daily. The tool is run on a machine with four Intel Xeon E5649 processors, each of which has 1 core, and 16.0 GB of RAM. The output from the RPO is the price recommendation for the merchants, which is also stored in the database for post-event margin analysis.

As the business and competitive landscape change over time, it is important to update the 100 regression trees stored in R that are used to predict demand. To do this, we have implemented an automated process that pulls historical data from Rue La La's ERP system and creates 100 new regression trees to use for future predictions in order to guarantee the tool's effectiveness.

4.2. Field Experiment

Being able to estimate the tool's impact prior to implementation was key in gaining buy-in and approval from Rue La La executives to use the pricing decision support tool to price first exposure styles. Although not modeled explicitly, Rue La La was particularly concerned that raising prices would decrease sales. One reason for this concern is that Rue La La wants to ensure that their customers find great value in the products on their site; offering prices too high could lower this

perceived value, and Rue La La could experience customer attrition. Another reason for their concern is that a decrease in sales corresponds to an increase in leftover inventory. Since the "flash sales" business model is one that relies on customers perceiving scarcity of each item, an increase in leftover inventory would need to be sold in additional events, which over time could diminish the customers' perception of scarcity.

Preliminary analysis of the pricing decision support tool on historical data suggested that, in fact, the model recommended price increases had little to no effect on sales quantity[5]. Motivated by this analysis, we wanted to design an experiment to test whether implementing model recommended price increases would decrease sales. Ideally, we would have liked to design a controlled experiment where some customers were offered prices recommended by the tool and others were not; due to potentially inducing negative customer reactions from such an experiment, we decided not to pursue this type of experiment. Instead, we developed and conducted a field experiment that took place from January through May of 2014 and satisfied Rue La La's business constraints.

Our goal for the field experiment was to address two questions: (*i*) would implementing model recommended price increases cause a decrease in sales quantity, and (*ii*) what impact would the recommended price increases have on revenue?

4.2.1. Experimental Design

In January 2014, we implemented the pricing decision support tool and began monitoring price recommendations on a daily basis. Recall that the lower and upper bounds on each style's range of possible prices were set such that the tool is configured to only recommend price increases (or no price change). Over the course of five months, we identified a test set of approximately 1,300 {event, subclass} combinations (i.e. sets of competing styles) where the tool recommended price increases for at least one style; this corresponds to approximately 6,000 styles with price increase recommendations. We assigned the set of styles in each {event, subclass} combination which the model recommended price increases to either a "treatment group" or "control group". The assignments to treatment vs. control groups were made for each {event, subclass} combination rather than for each style because optimization results are valid only when price increase recommendations are accepted for all competing styles in the event.

We were interested in answering the two questions posed above for styles in different price ranges, since price is an important feature in our model and, to a certain extent, is a good representation of different types of styles that Rue La La offers and associated customer segments. Thus we created five categories based on the legacy price of the style - Categories A, B, C, D, and E - and corresponding treatment and control groups for each of these five categories. All styles in the same category have a similar price, with Category A representing the lowest priced styles and Category E representing the highest priced styles. For styles in a treatment group, we accepted the model

Category	Median Predicted Sell-Through	Median Price	Median Discount off MSRP	Median Relative Price of Competing Styles
A	-1%	0%	0%	-7%
B	1%	0%	0%	-1%
C	-3%	0%	2%	0%
D	6%	-12%	2%	0%
E	0%	12%	0%	2%

Table 3 Percent increase in each treatment group's metric over the control group's metric

recommended price increases and raised prices accordingly. For styles in a control group, we did not accept the model recommended price increases and kept the legacy price (p_L). Denote p^* to be the actual price of the style that was offered to the customers. For styles in the treatment group, $p^* > p_L$, and for styles in the control group, $p^* = p_L$.

We wanted to ensure that we controlled for potential confounding variables which could bias the results of our experiment. Thus, as the field experiment progressed over the five-month period, we chose assignments of each {event, subclass} combination in order to ensure that the set of styles in the treatment vs. control groups for each category had a similar product mix in terms of the following metrics: median predicted sell-through ($\frac{\hat{d}_i}{C_i}$), median price, median discount off MSRP, and median relative price of competing styles. The price used to calculate all of these metrics (including the sales estimate $\hat{d}_i$) was p_L because we wanted the two groups to be as similar as possible before any price increases were applied. On a periodic basis, we would evaluate these metrics for each group in each category, and choose treatment vs. control group assignments in order to balance these metrics between the two groups; since Rue La La quickly adapts its assortment decisions to fashion trends throughout a selling season, we needed to be able to make these assignments throughout the field experiment as opposed to at the beginning of the experiment. A potential limitation of this assignment procedure is that it is not randomized and thus any differences observed between the treatment and control groups may be due to some other pre-existing attribute(s) of the styles in each group which we did not control for.

Table 3 shows the percent increase in each treatment group's metric over the control group's metric. For example, the percentages shown in the "Median Discount off MSRP" column are calculated as $\frac{\textit{median discount of treatment group styles} - \textit{median discount of control group styles}}{\textit{median discount of control group styles}}$. In most cases, the numbers in this table are close to zero, which implies that the medians in the two samples are the same. To test this more rigorously, we performed Mood's median test on each metric and in each category. Mood's median test is a special case of Pearson's χ^2 test which tests the null hypothesis that the population medians in the treatment and control groups are equal; see Corder and Foreman (2014) for more details on Mood's median test. Table 4 shows the results of this test for each category and metric.

Category	χ^2 Test Statistic for Predicted Sell-Through	χ^2 Test Statistic for Price	χ^2 Test Statistic for Discount off MSRP	χ^2 Test Statistic for Relative Price of Competing Styles	# Styles in Treatment Group	# Styles in Control Group
A	5.44	29.27**	0.00	33.76**	894	740
B	10.86*	29.16**	9.12*	9.12*	852	1086
C	11.16*	0.65	3.73	2.24	526	706
D	8.26*	4.34	3.31	0.00	418	509
E	0.25	2.00	3.17	1.10	278	209

* Reject null hypothesis at $\alpha = 0.05$. ** Reject null hypothesis at $\alpha = 0.01$.

Table 4 Results from Mood's median test

All but three of the tests we performed do not reject the null hypothesis at significance level $\alpha = 0.01$, suggesting that the medians of the treatment and control groups are the same. For Categories A and B, Mood's median test rejects the null hypothesis that the median price of the treatment and control groups are identical. In each of these two cases, the treatment group had significantly more styles whose price was less than the median price of the category's styles than the number of styles whose price was greater than the median. For Category A, Mood's median test also rejects the null hypothesis that the median relative price of competing styles of the treatment and control groups are identical. In this case, the treatment group had significantly more styles whose relative price was less than the median relative price of the category's styles than the number of styles whose relative price was greater than the median. Because of this, our results for Categories A and B may be biased, although the direction of such bias is uncertain. We believe that there could be a bias in overstating demand of the treatment group compared to the control group, since - all else equal - more styles with lower price should intuitively lead to higher demand.

Finally, it may be helpful to point out that the worst metrics in Table 3 (median price for Categories D and E) correspond to tests which did not reject the null hypothesis, whereas the null hypothesis is rejected for the median price metrics in Categories A and B which show no differences in Table 3. We note, however, that Table 3 compares the median price of the treatment group with the median price of the control group, whereas Table 4 compares the split of styles in each group below and above the category's median price.

Testing for Impact on Sell-Through

To estimate the impact of the price increases on sales quantity (d_i), we chose to focus on the actual sell-through ($\frac{d_i}{C_i}$) of the styles, which is effectively sales normalized by inventory. Intuitively, if price increases cause a decrease in sales, the sell-through of styles in the treatment group should be less than the sell-through of styles in the control group. For each of the five categories, we used the Wilcoxon rank sum test (also known as the Mann-Whitney test) to explore this rigorously. We chose to use this nonparametric test because it makes no assumption on the sell-through data

following a particular distributional form, and it is insensitive to outliers. Below are the three assumptions required for the test:

- *Assumption 1*: the styles in the treatment and control groups are random samples from the population.
- *Assumption 2*: there is independence within each group and between groups.
- *Assumption 3*: the metric being tested (e.g. $\frac{d_i}{C_i}$) is ordinal and can therefore be compared across all samples.

Recall from Table 3 that the difference in sell-through (and other price-related metrics) between the treatment and control groups are comparable, and thus we satisfy the first assumption. The independence in the second assumption comes from the fact that we assigned all styles in a given subclass and event to either a treatment or control group. The last assumption is the one that requires us to look at a metric such as sell-through as opposed to running a test on sales quantity, since sales quantity often depends on the amount of inventory available and thus is not comparable across styles.

We will describe the application of this test to our setting; for more details on the Wilcoxon rank sum test, see Rice (2006). Let F_s be the distribution of $\frac{d_i}{C_i}$ for styles given the legacy price (control), and let G_s be the distribution of $\frac{d_i}{C_i}$ for styles given the price increase (treatment). The null hypothesis (H_0) of our test is that raising prices has no effect on the distribution of $\frac{d_i}{C_i}$, i.e. $H_0 : G_s = F_s$, and this is compared to the one-sided alternative hypothesis (H_A) that raising prices decreases sell-through, i.e. $H_A : G_s < F_s$. To perform the test, we combined the sell-through data of all styles in the treatment and control groups, ordered the data, and assigned a rank to each observation. For example, the smallest sell-through would be assigned rank 1, the second smallest sell-through would be assigned rank 2, etc.; ties were broken by averaging the corresponding ranks. Then we summed the ranks of all of the treatment group observations. If this sum is statistically too low, the observations of $\frac{d_i}{C_i}$ in the treatment group are significantly less than the observations of $\frac{d_i}{C_i}$ in the control group and we reject the null hypothesis. Otherwise, we do not reject the null hypothesis, and any difference in observations of $\frac{d_i}{C_i}$ between the treatment and control groups is due to chance.

Testing for Impact on Revenue

To estimate the impact of model recommended price increases on revenue, we again used a Wilcoxon rank sum test on the same treatment and control groups. First we had to identify a metric that could be compared across different samples; for the same reason that d_i could not be used directly in the test described above (in order to satisfy the third assumption), we can not use revenue ($p^* d_i$) for this test. We define "available revenue" as the revenue that Rue La La could have earned if they had used the legacy price, i.e. $p_L C_i$, and the metric we used in our test is the "percent of available

revenue earned" which we define as $\frac{p^* d_i}{p_L C_i}$. Notice that the maximum percent of available revenue earned for a style in the control group is 100% because the actual revenue for these styles is $p_L d_i$, and the maximum possible quantity that can be sold is C_i. However, the maximum percent of available revenue earned for a style in the treatment group can be larger than 100% since $p^* > p_L$. A benefit of using this metric is that its predicted value using the legacy price is equivalent to sell-through ($\frac{p_L \hat{d}_i}{p_L C_i} = \frac{\hat{d}_i}{C_i}$), and thus the first assumption is satisfied for the same reason as above.

In contrast to our test of the impact on sell-through, we are interested in more than just identifying whether or not price increases cause an increase in the percent of available revenue earned. We would like to go one step further and quantify the impact on the percent of available revenue earned, and ultimately the impact on revenue. To do this, we need to make one more assumption:

- *Assumption 4*: the distribution functions of the percent of available revenue earned in the treatment and control groups are identical apart from a possible difference in location parameters.

We used the two-sample Kolmogorov-Smirnov test to test whether the predicted sell-through distributions were the same for the treatment and control groups, i.e. whether Assumption 4 holds. We found that Assumption 4 holds at significance level $\alpha = 0.1$ for Categories A, C, and D, $\alpha = 0.05$ for Category B, and $\alpha = 0.01$ for Category E.

Assumption 4 allows us to test a different null hypothesis which we can use to quantify the impact on revenue. Let F_r be the distribution of the percent of available revenue earned for styles in the control group, and let G_r be the distribution of the percent of available revenue earned for styles in the treatment group. The null hypothesis of our test is that raising prices has the effect of adding a constant (Δ) to what the percent of available revenue earned would have been if the legacy price was used, i.e. $H_0 : G_r(x) = F_r(x - \Delta)$; this is compared to the two-sided alternative $H_A : G_r(x) \neq F_r(x - \Delta)$. We used the Hodges-Lehmann estimator (HLΔ) for the Wilcoxon rank sum test to estimate the treatment effect Δ and constructed 90% and 95% confidence intervals of the possible values of Δ for which H_0 is not rejected. HLΔ is simply the median of all possible differences between a response in the treatment group and a response in the control group. The procedure we used to develop the nonparametric confidence interval around this estimate is described in Rice (2006) immediately following the description of the Wilcoxon rank sum test.

4.2.2. Results: Impact on Sell-Through

Table 5 presents the p-values of the Wilcoxon rank sum test for each of the five categories[6], as well as our decision on whether or not to reject the null hypothesis based on a significance level of $\alpha = 10\%$. Note that for Category A, we would choose to reject the null hypothesis even at the significance level $\alpha = 0.5\%$, whereas for Categories C-E, we would choose not to reject the null hypothesis even at significance levels higher than $\alpha = 30\%$. The large p-value of 0.842 for styles

Category	p-value	Result
A	0.004	Reject H_0
B	0.102	Do Not Reject H_0
C	0.347	Do Not Reject H_0
D	0.521	Do Not Reject H_0
E	0.842	Do Not Reject H_0

Table 5 Wilcoxon rank sum test results for the impact on sell-through

in Category E suggests that perhaps raising prices on these high price point styles may actually increase sales, although this is not a rigorous conclusion of the test.

These results show that raising prices on the lowest priced styles (styles in Category A) does have a negative impact on sell-through, whereas raising prices on the more expensive styles does not. We believe that there are two key factors contributing to these somewhat surprising results. First, since Rue La La offers very deep discounts, they may already be well below customers' reservation prices (i.e. customers' "willingness to pay"), such that small changes in price are still perceived as great deals. In these cases, sell-through is insensitive to price increases. Thus the pricing model finds opportunities to slightly increase prices without significantly affecting sell-through; consumers still benefit from a great sale, and Rue La La is able to maintain a healthy business.

A second key factor is that a change in demand corresponds to a smaller change in sales when limited sizes or inventory are available. As a simple example, consider the case where there is only 1 unit of inventory of a particular size. Demand for that size may be 10 units when priced at \$100 and only 2 units when priced at \$110, but *sales* would be 1 unit in both cases. Typically the higher priced styles have less inventory than the lower priced styles, which suggests that this factor may be contributing to the results we see.

Overall, the results to identify whether or not the model recommended price increases would decrease sales were promising: for most price points, the results of the tests imply that raising prices according to the model's recommendations do not decrease sales. This certainly helped mitigate concerns that Rue La La executives had over implementing the tool to price all first exposure styles. For the lowest priced styles, the maximum price increase of \$15 may be too large for consumers to still perceive a great deal, and this could be why we see a negative impact on sell-through. Rue La La has acted on this hypothesis and has limited price increases to \$5 for styles $p_L < \$50$.

4.2.3. Results: Impact on Revenue

For each of the five categories, Table 6 presents the Hodges-Lehmann estimator (HLΔ) for the *additive* increase in the percent of available revenue earned due to the model's recommended price increases, along with a 90% and 95% confidence interval of the estimate. As an example of how to interpret the numbers in the table, say that the percent of available revenue earned for a Category D style in the control group is 50%. HL$\Delta = 7.1\%$ means that if prices had been raised according

Category	HLΔ	90% Confidence Interval	95% Confidence Interval
A	-1.8%	[-6.4%, 3.3%]	[-7.6%, 4.5%]
B	5.5%	[1.8%, 9.6%]	[0.5%, 10.5%]
C	6.1%	[0.5%, 12.2%]	[-1.1%, 13.8%]
D	7.1%	[1.6%, 13.7%]	[0.0%, 15.1%]
E	14.9%	[2.7%, 31.7%]	[0.0%, 37.0%]
Overall	3.9%	[1.4%, 6.3%]	[0.7%, 7.0%]

Table 6 Estimate of additive increase in percent of available revenue earned due to raising prices

to the model's recommendations, the estimated percent of available revenue earned would increase to 57.1% with an associated 90% confidence interval of [51.6%, 63.7%].

For Category A, HL$\Delta < 0$ means that raising prices has actually decreased the percent of available revenue earned. This is not surprising given the results shown in Section 4.2.2 which strongly imply that raising prices decreases sell-through. The fact that HL$\Delta < 0$ further implies that the increase in per unit price is not enough to offset the decrease in sales quantity, thus overall decreasing the percent of available revenue earned. On a positive note, Categories B, C, D, and E all show HL$\Delta > 0$ with 90% confidence intervals containing only positive values of the estimates. This suggests that raising prices according to the model's recommendations increases the percent of available revenue earned for these categories. The estimated increase in Category E is particularly high at 14.9%, although with less data in this category relative to the other categories, the confidence interval's width is larger. With the exception of Categories A and C, even the more conservative 95% confidence intervals do not contain negative estimates of the additive percent increase of available revenue.

The percent of available revenue earned metric is not a common metric used at Rue La La, so we would like to convert this metric to an estimate of the percent increase in revenue, a more widely used metric in practice. Recall from Section 4.2.1 that we cannot measure the percent (or absolute) increase in revenue directly because we must satisfy the assumptions outlined for the hypothesis test. Thus, below are the steps we took to estimate the percent increase in revenue for a given value of Δ. Steps 1 and 2 essentially calculate the counterfactual revenue for each style had it been in the other group (treatment or control), and Step 3 then estimates each style's percent increase in revenue; this counterfactual analysis inherently assumes that the value of Δ impacts each style in a given category in the same way.

1. For styles in the control group, we estimated what the revenue would have been if the price had been increased according to the model's recommendation ("estimated treatment revenue") as $p_L d_i + \Delta p_L C_i$. We further bound this number below by \$0 and above by the maximum revenue that could have been achieved had the style been priced according to the model's recommendation.

2. For styles in the treatment group, we estimated what the revenue would have been if the price had been p_L ("estimated control revenue") as $p^* d_i - \Delta p_L C_i$. We further bound this number below

Category	Estimate of Percent Increase in Revenue	Estimate of 90% Confidence Interval	Estimate of 95% Confidence Interval
A	-3.4%	[-11.5%, 7.7%]	[-13.5%, 10.1%]
B	11.4%	[3.9%, 19.2%]	[1.1%, 21.0%]
C	12.5%	[1.1%, 23.4%]	[-2.0%, 26.6%]
D	13.7%	[3.4%, 22.8%]	[0.0%, 25.2%]
E	23.8%	[5.4%, 47.6%]	[0.0%, 56.7%]
Overall	9.7%	[2.3%, 17.8%]	[0.0%, 20.2%]

Table 7 Estimate of percent increase in revenue due to raising prices

by \$0 and above by the maximum revenue that could have been achieved had the style been priced at p_L.

3. For all styles, we divided the treatment revenue by the control revenue to find the percent increase in revenue. Note that for styles in the control group, the treatment revenue is estimated as in Step 1 whereas the control revenue is the actual revenue, and vice versa for styles in the treatment group. We omitted the $< 5\%$ of the styles whose control revenue was \$0.

For each of the values of Δ shown in Table 6 (i.e. $\Delta = HL\Delta$, the lower confidence interval bounds, and the upper confidence interval bounds), we used the above steps to estimate an associated percent increase in revenue for each style; taking the median within each category gives us the results shown in Table 7. As an example of how to interpret the numbers in the table, say that the revenue for a Category D style in the control group is \$500. If prices had been raised according to the model's recommendations, the estimated revenue would be $\$500(1.137) = \$568.50$ with an estimated 90% confidence interval of [\$517, \$614]. Note that the percentages shown in Table 7 are multiplicative, in contrast with the additive percentages shown in Table 6.

Similar insights are shown in this table. For Category A, we expect that revenue will likely decrease when prices are increased according to the model's recommendations; thus the per unit price increases do not make up for the decrease in sell-through. For Categories B, C, and D, we expect approximately an 11-14% increase in revenue due to raising prices. Category E shows the largest percent increase in revenue, although the confidence interval is much wider. Overall in our field experiment, we estimate a 9.7% increase in revenue with an associated 90% confidence interval of [2.3%, 17.8%] from using the model's price recommendations. Because of these positive results, we are now using the pricing decision support tool to make price recommendations on hundreds of new styles every day.

4.2.4. Source of Revenue Increases

We designed our field experiment to evaluate the impact of using our model's recommended prices vs. Rue La La's legacy prices (the "status quo"). Could most of these revenue gains be captured using a simpler technique without price optimization? What would be the estimated impact of only using demand forecasting as opposed to integrating forecasting with price optimization? These are

interesting questions that would best be answered by designing simpler forecasting models and/or pricing policies and performing additional field experiments. Since doing so was out of the scope of this project, we provide below a back-of-the-envelope analysis to help shed some light on the answers to these questions.

Consider the setting where Rue La La has access to style-level demand forecasts identical to those described in Section 2 for the case when all prices are set to p_L. Of course having only forecasts will not impact revenue unless prices are changed as a result of these forecasts. Thus we must specify how Rue La La would use demand forecasts to adjust the prices of its styles without using optimization. We discussed this hypothetical situation with our main contacts at Rue La La in order to determine how they would use this forecast information to make better pricing decisions. The consensus was that they would likely choose to raise prices on styles that were predicted to sell out, in hopes that they would earn more revenue with little impact on sell-through. The determination of how much to raise the price for each style would be done on a style-by-style basis with upper bounds on price increases as described in Section 4.1.

To estimate the impact of raising prices only on styles that were predicted to sell out, we simply looked at the styles in our field experiment that were predicted to sell out at price p_L, and we attributed the revenue increase from these styles to having a better forecast. Specifically, we used the same procedure as described in Section 4.2.3 to convert the results from Table 6 to an estimate of the percent increase in revenue, but this time we only considered styles that were predicted to sell out; we estimated that raising prices on this subset of styles resulted in approximately an 11.0% increase in revenue with an associated 90% confidence interval of [2.7%, 16.7%]. Furthermore, by dividing the total estimated treatment revenue for all styles which were expected to sell out by the total estimated treatment revenue for all styles, we found that approximately 30% of the estimated increase in (absolute) revenue in the field experiment can be attributed to styles that were predicted to sell out. In other words, we expect that Rue La La could have achieved approximately 30% of the model's benefit simply by using our demand forecasts.

We would like to highlight that this analysis is only meant to provide a back-of-the-envelope estimate of the impact of forecasting vs. integrating forecasting with price optimization, primarily due to two key issues. First, the amount each style's price was raised was due to the output of the price optimization tool and is dependent on the other competing styles in the event. Prices may or may not have been raised by the same amount had the merchants only been given demand forecasts for p_L. Second, raising prices on these styles likely has some effect on the sales of other competing styles in the event, and we are not incorporating such an effect in our analysis. Similarly, sales of the styles that were predicted to sell out may have been affected by price changes of other competing styles in the field experiment. Despite these issues, we believe our method provides

a rough estimate of the portion of the revenue increases that would have been achieved simply by using demand forecasts, and it suggests that a majority of the benefit of the pricing decision support tool comes from the integration of demand forecasting with price optimization.

5. Conclusion

In this paper we shared our work with Rue La La on the development and implementation of a pricing decision support tool used to maximize first exposure styles' revenue. One of the development challenges was predicting demand for items that had never been sold before. We found that regression trees with bagging outperformed other regression methods we tested on a variety of performance metrics. Unfortunately, their nonparametric structure - along with the fact that each style's demand depends upon the price of all competing styles - led to a seemingly intractable price optimization problem. We developed a novel reformulation of the optimization problem and created an efficient algorithm to solve this problem on a daily basis to price the next day's first exposure styles. We conducted a field experiment to evaluate our pricing decision support tool and showed an expected increase in first exposure styles' revenue in the test group of approximately 9.7% with an associated 90% confidence interval of [2.3%, 17.8%], while minimally impacting aggregate sales. These positive results led to the recent adoption of our pricing decision support tool for daily use.

There are several key takeaways from our research for both practitioners and academics. First, we developed an efficient algorithm to solve a multi-product price optimization model that incorporates reference price effects which can be used by other retailers to set prices of new products. This extends beyond the flash sales setting and can be used by retailers who make production/purchasing decisions well before the selling season begins, and whose forecast accuracy for a given style is likely to improve as the beginning of the selling season approaches. As soon as production/purchasing decisions have been made, the cost of this inventory can be considered a sunk cost; as the selling season approaches, the retailer can improve his demand forecasts and then use our optimization model to set prices.

Another key takeaway is that we showed how combining machine learning and optimization techniques into a pricing decision support tool has made a substantial financial impact on Rue La La's business. We hope that the success of this pricing decision support tool motivates retailers to investigate similar techniques to help set initial prices of new items, and, more broadly, that researchers and practitioners will use a combination of machine learning and optimization to harness their data and use it to improve business processes.

Finally, we encourage further exploration of using nonparametric regression techniques to predict demand. Even extending beyond pricing, predicting demand accurately is a necessary requirement for input into many operations problems, and thus we challenge researchers and practitioners to

explore new - and possibly less structured - demand prediction models. In particular, we believe that regression trees would be effective in predicting demand for (*i*) new products and (*ii*) products whose price can be considered a signal of quality.

This paper would not be complete if we did not mention a few directions of potential future work. Recently, we have begun working with Rue La La on a project to help them identify how much additional revenue could be earned if prices were allowed to change throughout the course of the event. With this goal in mind, we have developed a dynamic pricing algorithm that learns a customer's purchase probability of a product at each price point by observing real-time customer purchase decisions, and then uses this knowledge to dynamically change prices to maximize total revenue throughout the event. Our algorithm builds upon the well-known Thompson Sampling algorithm used for multi-armed bandit problems by creatively incorporating inventory constraints into the model and algorithm. We show that our algorithm has both strong theoretical performance guarantees as well as promising numerical performance results when compared to other algorithms developed for the same setting. See Ferreira et al. (2015) for more details on this work.

Another direction that we have begun pursuing centers around quantifying the benefit of the flash sales business environment. In particular, we aim to identify the impact of frequent assortment rotations on sales. Consider, for example, a retailer that would like to sell 10 similar items over the course of a selling season. Traditionally, a retailer would offer all 10 products concurrently to the customer; in the flash sales environment, the retailer may offer only one product to the customer at a time, for 10 disjoint periods throughout the selling season. In the first case, the consumer is able to observe all 10 products before selecting her favorite ones to buy. However in the flash sales environment, the consumer must decide whether or not to buy each item before viewing the remaining items that will be sold in the season; if she chooses not to buy, she will not have the opportunity to buy that product at a later time. We aim to (*i*) identify types of products where frequent assortment rotation in a flash sales environment would lead to an increase in total retail sales, and (*ii*) quantify this benefit (see Ferreira and Simchi-Levi (2015)).

Our collaboration with Rue La La has shed light on the unique challenges present in the relatively new and growing flash sales industry. As this work illustrates, there is potential for academics and practitioners to work together to develop new operations management models and techniques tailored to this industry, and ultimately guide the industry's future growth.

Endnotes

1. In some cases, the contract is such that the designer commits to selling up to X units of an item to Rue La La in a given time window, but Rue La La is not committed to purchasing anything. Rue La La plans an event within the time window, receives customer orders up to X units, and

then purchases the quantity it has sold. There are a few changes to the model and implementation steps due to this type of contract, but for ease of exposition they have been left out of the paper.

2. Returns re-enter the process flow at this point and are treated as remaining inventory.

3. Detailed descriptions of these models can be found in Hastie et al. (2009) (least squares, principal components, partial least squares, regression trees) and Talluri and Van Ryzin (2005) (multiplicative, semilogarithmic).

4. Another common way to address the issue of overfitting in regression trees is by using random forests. We chose to use bagging instead of random forests for better interpretability.

5. A discussion of this historical analysis has been left out for brevity.

6. We used the approximation presented below Table 8 of Appendix B in Rice (2006) to calculate the critical values.

Acknowledgments

We thank Murali Narayanaswamy, the Vice President of Pricing & Operations Strategy at Rue La La, Jonathan Waggoner, the Chief Operating Officer at Rue La La, and Philip Roizin, the Chief Financial Officer at Rue La La, for their continuing support, sharing valuable business expertise through numerous discussions, and providing us with a considerable amount of time and resources to ensure a successful project. The integration of our pricing decision support tool with their ERP system could not have been done without the help of Hemant Pariawala and Debadatta Mohanty. We also thank the numerous other Rue La La executives and employees for their assistance and support throughout our project. This research also benefitted from discussions with Roy Welsch (MIT), Özalp Özer (UT Dallas), Matt O'Kane (Accenture), Andy Fano (Accenture), Paul Mahler (Accenture), Marjan Baghaie (Accenture), and students in David Simchi-Levi's research group at MIT. Finally, we thank the referees and area editor, whose comments significantly helped the presentation and analysis in this paper. This work was supported by Accenture through the MIT Alliance in Business Analytics.

References

Anupindi, R., M. Dada, S. Gupta. 1998. Estimation of Consumer Demand with Stock-Out Based Substitution: An Application to Vending Machine Products. *Marketing Science* **17**(4) 406–423.

Berry, S., J. Levinsohn, A. Pakes. 1995. Automobile Prices in Market Equilibrium. *Econometrica* **63**(4) 841–890.

Birge, J., J. Drogosz, I. Duenyas. 1998. Setting Single-Period Optimal Capacity Levels and Prices for Substitutable Products. *The International Journal of Flexible Manufacturing Systems* **10** 407–430.

Bitran, G., R. Caldentey. 2003. An Overview of Pricing Models for Revenue Management. *Manufacturing & Service Operations Management* **5**(3) 203–229.

Caro, F., J. Gallien. 2012. Clearance Pricing Optimization for a Fast-Fashion Retailer. *Operations Research* **60**(6) 1404–1422.

Choi, T. M. 2007. Pre-season Stocking and Pricing Decisions for Fashion Retailers with Multiple Information Updating. *International Journal of Production Economics* **106**(1) 146–170.

Corder, G. W., D. I. Foreman. 2014. *Nonparametric Statistics: A Step-by-Step Approach*. John Wiley & Sons.

Şen, A. 2008. The US Fashion Industry: A Supply Chain Review. *International Journal of Production Economics* **114**(2) 571–593.

Du, D., P. Pardalos. 1998. *Handbook of Combinatorial Optimization*, vol. 1. Kluwer Academic Publishers.

Elmaghraby, W., P. Keskinocak. 2003. Dynamic Pricing in the Presence of Inventory Considerations: Research Overview, Current Practices, and Future Directions. *Management Science* **49**(10) 1287–1309.

Emery, F. 1970. Some Psychological Aspects of Price. B. Taylor, G. Wills, eds., *Pricing Strategy*. Brandon/Systems Press, Princeton, N.J., 98–111.

Everitt, B., S. Landau, M. Leese, D. Stahl. 2011. *Cluster Analysis*. 5th ed. Wiley.

Ferreira, K. J., D. Simchi-Levi. 2015. Choosing an Assortment Rotation Strategy to Boost Sales. *Working Paper* .

Ferreira, K. J., D. Simchi-Levi, H. Wang. 2015. Online Network Revenue Management Using Thompson Sampling. *Working Paper* .

Garro, A. 2011. New Product Demand Forecasting and Distribution Optimization: A Case Study at Zara. *Doctoral dissertation, Massachusetts Institute of Technology* .

Gaur, V., M. L. Fisher. 2005. In-Store Experiments to Determine the Impact of Price on Sales. *Production and Operations Management* **14**(4) 377–387.

Hastie, T., R. Tibshirani, J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer.

Lau, L. C., R. Ravi, M. Singh. 2011. *Iterative Methods in Combinatorial Optimization*. Cambridge University Press.

Levy, M., D. Grewal, P. K. Kopalle, J. D. Hess. 2004. Emerging Trends in Retail Pricing Practice: Implications for Research. *Journal of Retailing* **80**(3) xiii–xxi.

Little, J., J. Shapiro. 1980. A Theory for Pricing Nonfeatured Products in Supermarkets. *The Journal of Business* **53**(3) S199–S209.

LON, Local Offer Network. 2011. The Daily Deal Phenomenon: A Year in Review.

Maddah, B., E. Bish. 2007. Joint Pricing, Assortment, and Inventory Decisions for a Retailers Product Line. *Naval Research Logistics (NRL)* **54**(3) 315–330.

Mazumdar, T., S. P. Raj, I. Sinha. 2005. Reference Price Research: Review and Propositions. *Journal of Marketing* **69**(4) 84–102.

McKitterick, W. 2015. Online Fashion Sample Sales in the US. Tech. rep., IBISWorld Industry Report OD5438.

Musalem, A., M. Olivares, E. T. Bradlow, C. Terwiesch, D. Corsten. 2010. Structural Estimation of the Effect of Out-of-Stocks. *Management Science* **56**(7) 1180–1197.

Natter, M., T. Reutterer, A. Mild, A. Taudes. 2007. Practice Prize Report - An Assortmentwide Decision-Support System for Dynamic Pricing and Promotion Planning in DIY Retailing. *Marketing Science* **26**(4) 576–583.

Ostapenko, N. 2013. Online Discount Luxury: In Search of Guilty Customers. *International Journal of Business and Social Research* **3**(2) 60–68.

Özer, Ö., R. Phillips, eds. 2012. *The Oxford Handbook of Pricing Management*. Oxford University Press.

Reibstein, D., H. Gatignon. 1984. Optimal Product Line Pricing: The Influence of Elasticities and Cross-Elasticities. *The Journal of Business* **21**(3) 259–267.

Rice, J., ed. 2006. *Mathematical Statistics and Data Analysis*. Cengage Learning.

Smith, S., D. Achabal. 1998. Clearance Pricing and Inventory Policies for Retail Chains. *Management Science* **44**(3) 285–300.

Strobl, C., A. L. Boulesteix, T. Kneib, T. Augustin, A. Zeileis. 2008. Conditional Variable Importance for Random Forests. *BMC Bioinformatics* **9**(1) 307.

Subrahmanyan, S. 2000. Using Quantitative Models for Setting Retail Prices. *Journal of Product & Brand Management* **9**(5) 304–320.

Talluri, K. T., G. J. Van Ryzin. 2005. *The Theory and Practice of Revenue Management*. Springer.

Vulcano, G., G. Van Ryzin, R. Ratliff. 2012. Estimating Primary Demand for Substitutable Products from Sales Transaction Data. *Operations Research* **60**(2) 313–334.

Winer, R. 1985. A Price Vector Model of Demand for Consumer Durables: Preliminary Developments. *Marketing Science* **4**(1) 74–90.

Wolverson, R. 2012. High and Low: Online Flash Sales Go Beyond Fashion to Survive. *Time Magazine* **180**(19) Special Section 9–12.

Wu, J., L. Li, L. D. Xu. 2014. A Randomized Pricing Decision Support System in Electronic Commerce. *Decision Support Systems* **58** 43–52.

REGRESSOR	DESCRIPTION
Color Popularity	Historically, the percent of items sold in the style's department that were the same color.
Size Popularity	The percent of the size curve (described in Section 2.3) covered by sizes available for the style. This represents the breadth of sizes available to customers.
Brand Type A/B	Rue La La has an internal categorization of brands (based on qualities like, e.g., historical sales volume of the brand); we assign binary variables denoting whether or not a style's brand is in category A or B.
Brand Popularity	We included several variables to quantify a brand's "popularity", including: • Brand Popularity by Revenue: Historically, if the average revenue this brand earned in branded events is greater than a certain threshold percent more than the overall average revenue generated in branded events, this binary variable is set to 1 (0 otherwise). • Brand Popularity by Sell-Through: Similar to Brand Popularity by Revenue, except based on sell-through as opposed to revenue. • Brand MSRP Index: Historically, the average MSRP over all items sold in this subclass and for this brand, divided by the average MSRP over all items sold in this subclass. Averages are weighted by quantity sold, and the default value is 1. This represents the brand's price position in that subclass.
Event Type	Events are classified as either branded (all styles in event are from a single brand) or curated (styles from multiple brands are sold). This is a binary variable representing whether or not the event is branded.

Figure 10 Description of select features

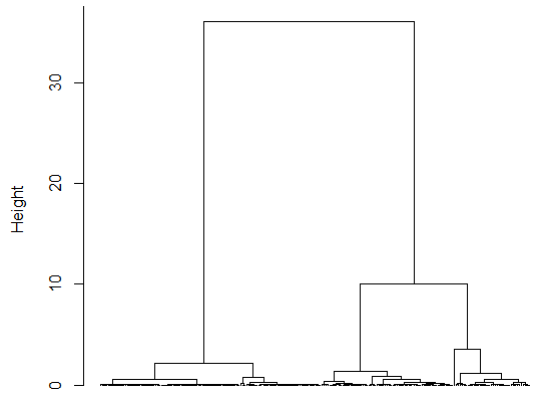

Figure 11 Cluster dendrogram for 2-day events

Appendix A: Description of Features

Figure 10 provides a description for the less intuitive features shown in Figure 5.

Appendix B: Demand Estimation - Clustering and Evaluation

For each event length, we first plotted a dendrogram using Ward's minimum variance method in hierarchical clustering to get an idea as to how many demand curves in which we should aggregate our data. Figure 11 shows the dendrogram for all the demand curves with a 2-day event length. The height of each vertical bar represents the increase in total within-cluster variance upon agglomeration of the two sets of demand curves, and each tick on the x-axis represents a demand curve associated with a different set of factors.

Looking at the dendrogram, it is clear that we would like 3-4 clusters because there is substantial benefit in terms of decreasing within-cluster variance when separating the factors into 3 or 4 groups. We split the set of factors by the clusters they were assigned (both for 3 and 4 clusters) and then tried to identify commonalities between each set of factors that would give us insight into the main factors that drive differences in demand curves. Overwhelmingly for 2-day events, the 3 clusters were primarily separated by start time of day - 11:00am, 3:00pm, and 8:00pm - and adding a 4th cluster seemed to further separate events with weekend morning start times. The same analysis was done for the other possible event lengths and similar insights

were made; for 4-day events, a couple of departments warranted their own demand curves, too. To test the robustness of our clustering technique, we performed a similar analysis using k-means clustering and found that the same factors drive the separation of clusters.

To evaluate the accuracy of our method to "unconstrain" demand to estimate sales (i.e. our "demand unconstraining" method), we gathered hourly sales data of each first exposure SKU sold in an event during the 9-month period after the demand curves had been created; this ensures that we are not testing on the same data that was used to create the clusters. We focus only on those items that did not sell out, i.e. when $u_{is} = d_{is}$. For each SKU, every hour that there was at least one unit sold, we considered that to be the hour the SKU stocked out, and we estimated the total demand of the SKU using our demand curves, $\tilde{u}_{is}$. We then compared this to $u_{is} = d_{is}$ to evaluate our demand unconstraining method's performance.

For example, say 3 units of a SKU were sold in a two-day event - one in the first hour and two in the third hour. Note that for this SKU, $u_{is} = d_{is} = 3$. First, we consider the hypothetical case where the item only has one unit in inventory and thus sold out in the first hour. We use the appropriate demand curve (depending on event start time of day and day of week) to estimate total sales, say $\tilde{u}_{is} = 5$. Second, we consider the hypothetical case where the item only has 3 units in inventory and thus sold out in the third hour. Again, we use the appropriate demand curve to estimate total sales, say $\tilde{u}_{is} = 6$. Note that our estimates need not be the same as they are based on two different hypothetical situations.

We followed the above steps for each SKU and calculated the error $(\tilde{u}_{is} - u_{is})$ for each of our estimations. We compared our approach ("demand unconstraining") to the approach where potential lost sales are not considered ("zero lost sales"), i.e. when actual sales are used as the estimate for demand. Figure 12 presents the hourly mean absolute error for 2-day events and SKUs in Rue La La's top 5 departments using our demand unconstraining approach vs. not accounting for lost sales; the data consists of over 50,000 SKUs with an average inventory of 9.7 units per SKU. The graph on the left shows the hourly mean absolute errors for SKUs with less than 10 units sold, whereas the graph on the right shows the same metrics for SKUs with at least 10 units sold.

The mean absolute error is very small for SKUs with less than 10 units sold and there is a negligible difference in this metric between our demand unconstraining approach and the approach that does not consider potential lost sales. With that said, it is important to recognize that the latter approach *always* under-estimates demand, which would lead to a systematic under-estimation of true demand; our proposed method using the demand curves sometimes over-estimates demand and sometimes under-estimates demand. As shown in the graph on the right, for SKUs with at least 10 units sold, our demand unconstraining approach significantly outperforms the approach that does not consider potential lost sales. As expected in both graphs, the error decreases with the time into the event that the stockout occurs since there are fewer remaining hours with potential lost sales. Overall, our demand unconstraining approach for estimating lost sales appears to work well for Rue La La.

One concern with our methodology of estimating lost sales is that there is some - potentially unobservable - factor(s) that is different between the set of SKUs that stocked out vs. the set of SKUs that did not stock out. If this is the case, then building the demand curves using only the data for SKUs that did not stock out and applying them to SKUs that did stock out could be problematic. With that said, it appears that the

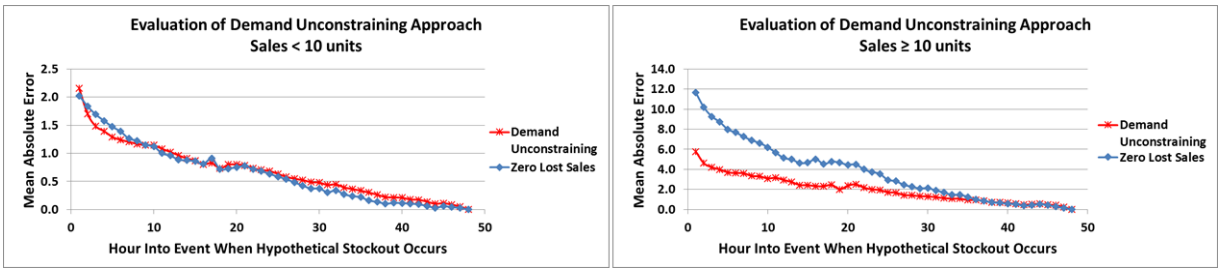

Figure 12 **Performance evaluation of demand unconstraining approach**

Total number of styles	3,071
Total number of SKUs	14,148
Average price	\$90
Average units sold per style	31
Average demand per style	39
Average inventory per style	47

Table 8 **Descriptive statistics of the test data set for Rue La La's largest department**

	Coefficient of Determination (R^2)	Median Absolute Error (MEDAE)	Median Absolute Percentage Error (MEDAPE)	Median Absolute Sell-Through Error (MEDASTE)
Demand	$1-\frac{\sum_{i\in\mathcal{N}}(\hat{u}_i-u_i)^2}{\sum_{i\in\mathcal{N}}(u_i-\bar{u})^2}$	$\text{median}(\lvert\hat{u}_i-u_i\rvert)$	$\text{median}\left(\frac{\lvert\hat{u}_i-u_i\rvert}{u_i}\right)$	-
Sales	$1-\frac{\sum_{i\in\mathcal{N}}(\hat{d}_i-d_i)^2}{\sum_{i\in\mathcal{N}}(d_i-\bar{d})^2}$	$\text{median}(\lvert\hat{d}_i-d_i\rvert)$	$\text{median}\left(\frac{\lvert\hat{d}_i-d_i\rvert}{d_i}\right)$	$\text{median}\left(\frac{\lvert\hat{d}_i-d_i\rvert}{C_i}\right)$

Table 9 **Performance metrics used to compare regression models**

main difference between the set of SKUs that stocked out vs. did not stock out is the amount of inventory available for each SKU, rather than product or event characteristics. For the data set used in these results, the average inventory of SKUs that did not stock out was 9.7 units; for SKUs that did stock out (in the top 5 departments and sold in two-day events in the same time period) the average inventory was only 3.4 units.

Appendix C: Demand Prediction Model Accuracy

Recall that we originally randomly divided our data into training and testing data sets, and the regression models were built using the training data. Here we apply the regression models to the testing data and describe why we chose regression trees as the foundation of our demand prediction model. We report only these detailed results for Rue La La's largest department (by revenue); the results for other departments are similar and do not add much to the discussion. Table 8 presents summary statistics of our testing data set.

We used the metrics listed in Table 9 to compare regression models. Recall that d_i is the actual sales of style i, and u_i is the demand for style i estimated as described in Section 2.2. Define $\bar{d}=\frac{1}{n}\sum_{i\in\mathcal{N}}d_i$ and $\bar{u}=\frac{1}{n}\sum_{i\in\mathcal{N}}u_i$, where n is the number of styles in the test data. Recall that $\hat{u}_i$ represents the predicted demand for style i, and $\hat{d}_i$ represents its predicted sales. Approximately 1% of styles had $d_i=0$, and these styles were removed from the analysis of the MEDAPE metric.

Note that three of the metrics are calculated both for the demand predictions and sales predictions; since sell-through depends directly on inventory, we only calculated MEDASTE for sales. Although the regression

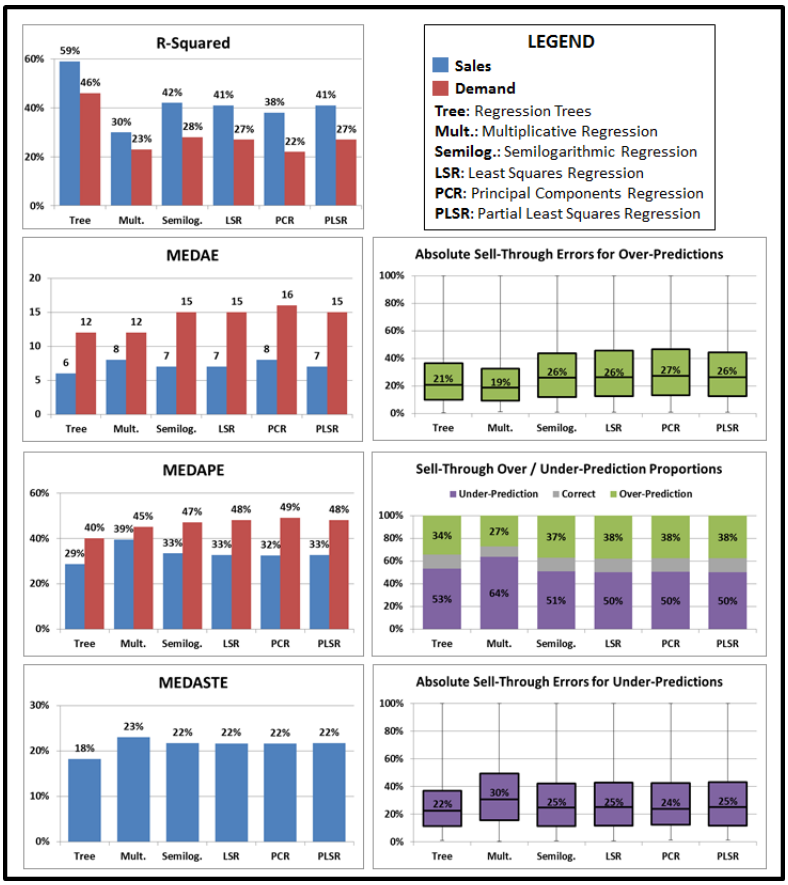

Figure 13 **Performance results comparison of regression models**

models are built to predict u_i, we believe that analyzing the prediction accuracy of the overall demand prediction model's output, d_i, is more important since this is the input to our price optimization model. Intuitively, we are less concerned about prediction errors when both $u_{is} > C_{is}$ and $\hat{u}_{is} > C_{is}$ because in this case, the model predicts the item will stock out which indeed happened, i.e. $d_{is} = \hat{d}_{is}$.

Figure 13 shows a summary of the performance results for each of the six regression models tested. On the left-hand side, the results of the seven metrics defined in Table 9 are shown. On the right-hand side, we show further details on the sell-through predictions since sell-through is a very important metric for Rue La La. As is clearly shown in Figure 13, regression trees outperform the other models tested with respect to all seven metrics and thus were chosen for our demand prediction model.

Interestingly, as shown on the right-hand side of Figure 13, the proportion of under-predictions of d_i is between 50-53% for all models (with the exception of the multiplicative model), whereas for over-predictions of d_i it is only 34-38%. We think a main reason for this is our conservative use of the size curves. The inherent assumption made when applying these size curves is that sizes are not substitutable; although we took measures to ensure this assumption is generally valid (by mapping substitutable sizes to a single size), there are still some consumers who may buy a neighboring size if their size is unavailable. Assuming sizes are not substitutable leads to a more conservative demand prediction model because it assumes that all demand is lost for a size if it is unavailable.

We recognize that mean absolute error (MAE) and mean absolute percentage error (MAPE) are more commonly used than their median counterparts MEDAE and MEDAPE, respectively. One concern with

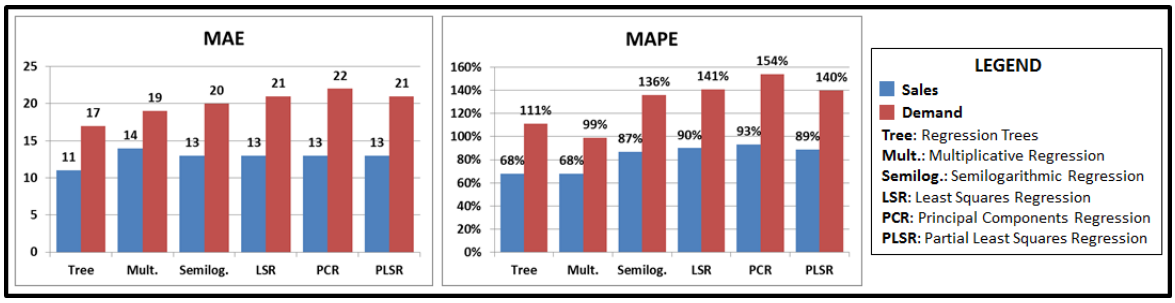

Figure 14 **Comparison of MAE and MAPE**

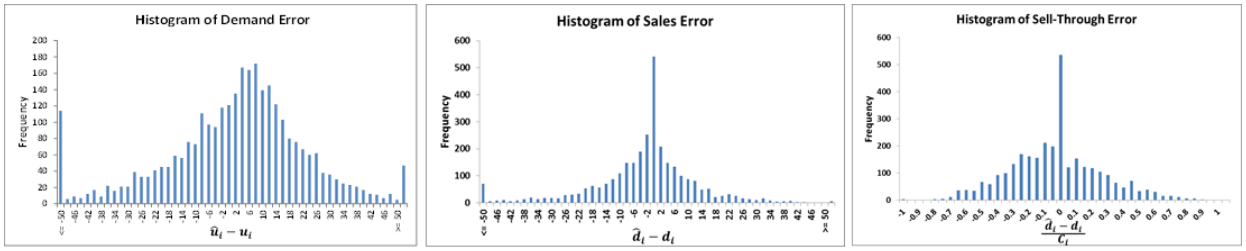

Figure 15 **Regression tree performance results**

using MAE and MAPE, though, is that they are not robust to outliers. The median provides a more robust measurement, especially for these right-skewed distributions. The absolute error for each style has a right-skewed distribution due to very high values of actual quantity sold for a few styles, whereas the absolute percentage error for each style has a right-skewed distribution due to very low values of actual quantity sold. For example, if $u_i = 1$ and $\hat{u}_i = 4$, style i's absolute percentage error is 300%; however in practice, this may be considered a very good prediction. Regardless, we evaluated MAE and MAPE for all regression models tested, and the results are shown in Figure 14. Again, regression trees often outperform the other methods.

Since we chose to use regression trees in our demand prediction model, we show more detailed regression tree performance results in Figure 15. The graph on the left is a histogram of the demand error, $\hat{u}_i - u_i$. The middle one shows the sales error, $\hat{d}_i - d_i$. The third histogram is for sell-through error, $\frac{\hat{d}_i - d_i}{C_i}$. Each histogram follows an approximately bell-shaped curve centered at zero; as expected from the above discussion, the histograms of $\hat{d}_i - d_i$ and $\frac{\hat{d}_i - d_i}{C_i}$ are skewed a bit to the left, likely due to our conservative application of size curves. In addition, it is clear from comparing the histograms of $\hat{u}_i - u_i$ and $\hat{d}_i - d_i$ that applying inventory constraints greatly reduces the error. This is due to the fact that all errors caused by items with either $u_{is} > C_{is}$ or $\hat{u}_{is} > C_{is}$ are mitigated in the sales prediction due to applying inventory constraints.

Finally, we calculated the correlation between our error term, $\hat{d}_i - d_i$, and all of the features in our regression as one test to identify if there were any systematic biases from potentially unobserved factors or endogeneity issues. The largest correlation (in magnitude) was approximately -0.11, suggesting that our application of regression trees may not suffer from these potential biases. For another test to try to identify potential omitted-variable bias, we identified four features with low variable importance. We removed each one from the feature set and built 100 regressions trees (used in bagging) without these features. We estimated demand on our test set and compared our demand estimates to those which were made using all features. The mean absolute error was only 1.5 units, and the mean absolute percentage error was 4.3%. It appears that our

predictions are not substantially changed when removing features with low variable importance, and thus we are likely not ignoring potentially unobserved features in the model.

Appendix D: Proofs of Theorems

Proof of Theorem 1

For every $k \in \mathcal{K}$, (LP_k) is a standard form linear optimization problem. By construction of the set $\mathcal{K}$, in particular that each $k \in \mathcal{K}$ is by definition a sum of possible prices for each style, we know that a feasible solution must exist and the feasible set is nonempty. Therefore by classical linear programming theory, we know that there exists at least one basic feasible solution. Furthermore, the constraint set is bounded and thus the optimal objective value can not be ∞, and there must exist a basic feasible solution which is optimal.

There are $N+1$ constraints (not including upper bound and non-negativity constraints) and NM variables in (LP_k); define matrix $\mathbf{A}$ to be the $(N+1) \times (NM)$ constraint coefficient matrix. With the exception of the uninteresting case where $M = 1$, $\mathbf{A}$ has linearly independent rows. Again from linear programming theory, we know that all non-basic variables in a basic feasible solution must be either 0 or 1. Since there are $NM - (N+1)$ non-basic variables, we know that at most $N+1$ variables can be fractional.

Combining the facts above, we know that an optimal solution to (LP_k) exists with no more than $N+1$ fractional variables. Consider now such an optimal solution to (LP_k) with no more than $N+1$ fractional variables, $x^*_{i,j}$ $\forall\ i \in \mathcal{N}, j \in \mathcal{M}$. In order to satisfy $\sum_{j \in \mathcal{M}} x_{i',j} = 1$ for some style i', it is impossible to do so with just one fractional variable associated with style i'. This implies that at most $\lfloor \frac{N+1}{2} \rfloor$ of the first N constraints of (LP_k) are satisfied with fractional variables. Equivalently, at most $\lfloor \frac{N+1}{2} \rfloor$ styles can have associated fractional variables; all of the other styles must be assigned exactly one price.

In this optimal solution, define $\mathcal{F} = \{i \mid x^*_{i,j} \neq \{0,1\}\ for\ some\ j\}$; in other words, $\mathcal{F}$ is the set of all styles that have associated fractional variables. Thus (LP_k) can be reduced to

$$\begin{aligned}
\sum_{i \notin \mathcal{F}} \sum_{j \in \mathcal{M}} p_j \tilde{D}_{i,j,k} x^*_{i,j} + \max \sum_{i \in \mathcal{F}} \sum_{j \in \mathcal{M}} p_j \tilde{D}_{i,j,k} x_{i,j} \\
st. \quad \sum_{j \in \mathcal{M}} x_{i,j} = 1 \quad \forall\ i \in \mathcal{F} \\
\sum_{i \in \mathcal{F}} \sum_{j \in \mathcal{M}} p_j x_{i,j} = k - \sum_{i \notin \mathcal{F}} \sum_{j \in \mathcal{M}} p_j x^*_{i,j} \\
0 \leq x_{i,j} \leq 1 \quad \forall\ i \in \mathcal{F},\ j \in \mathcal{M}
\end{aligned}$$

Let $\hat{k} \triangleq k - \sum_{i \notin \mathcal{F}} \sum_{j \in \mathcal{M}} p_j x^*_{i,j}$. Observe that the set of possible values of $\hat{k}$ has the same structure as $\mathcal{K}$, for a problem with $|\mathcal{F}|$ styles. In particular, the set of possible values of $\hat{k}$ lies within the range $|\mathcal{F}| * \min_{j \in \mathcal{M}}\{p_j\}$ to $|\mathcal{F}| * \max_{j \in \mathcal{M}}\{p_j\}$ in increments of 5. With this observation, we can see that this reduced optimization problem is identical to $(LP_{\hat{k}})$ with only $|\mathcal{F}| \leq \lfloor \frac{N+1}{2} \rfloor$ styles. As before, we know that there exists an optimal basic feasible solution to $(LP_{\hat{k}})$ with at most $\lfloor \frac{|\mathcal{F}|+1}{2} \rfloor$ of the styles having associated fractional variables. Thus, we can replace $x^*_{i,j}$ $\forall\ i \in \mathcal{F}, j \in \mathcal{M}$ with this new optimal basic feasible solution and achieve the same optimal objective value of (LP_k) with a basic feasible solution that has at most $\lfloor \frac{|\mathcal{F}|+1}{2} \rfloor$ styles with associated fractional variables.

Following the same technique as above, the optimal basic feasible solution to $(LP_{\hat{k}})$ can be separated into variables associated with two sets of styles - styles assigned a single price and styles with a fractional

assignment of prices - and similar steps to those above can be followed to find an optimal solution to $(LP_{\hat{k}})$ that has even fewer styles with a fractional assignment of prices. Since $\lfloor \frac{N+1}{2} \rfloor < N$ when $N \geq 2$, as long as there are at least 2 styles with a fractional assignment of prices, each iteration of this technique will reduce the number of styles with a fractional assignment of prices until there is at most only one style remaining that is not assigned a single price. If every style remaining is assigned a single price, we have an optimal solution to (IP_k), and $z^*_{LP_k} = z^*_{IP_k}$. Otherwise, denote the single style remaining with fractional price assignments as style $\hat{i}$. The final iteration of the above steps results in the following optimization problem:

$$\begin{aligned} \max & \sum_{j \in \mathcal{M}} p_j \tilde{D}_{\hat{i},j,k} x_{\hat{i},j} \\ st. \quad & \sum_{j \in \mathcal{M}} x_{\hat{i},j} = 1 \\ & \sum_{j \in \mathcal{M}} p_j x_{\hat{i},j} = \hat{k} \\ & 0 \leq x_{\hat{i},j} \leq 1 \quad \forall\, j \in \mathcal{M} \end{aligned}$$

There are only two constraints in this problem, and therefore there exists an optimal basic feasible solution with exactly 2 fractional variables, given the case where an integer optimal solution does not exist. We can construct a feasible integer solution to (LP_k) which only differs from the constructed optimal solution to (LP_k) by the variables associated with style $\hat{i}$. To do so, simply set $x_{\hat{i},j} = 1$ if $p_j = \hat{k}$ and $x_{\hat{i},j} = 0$ otherwise. In this case, the possible values of $\hat{k}$ lie within the range $\min_{j \in \mathcal{M}}\{p_j\}$ to $\max_{j \in \mathcal{M}}\{p_j\}$ in increments of 5, so we are certain that there exists a j such that $p_j = \hat{k}$.

This feasible integer solution to (LP_k) is clearly a feasible integer solution to (IP_k) which provides a lower bound on the optimal solution to (IP_k). The optimal objective value of this feasible solution only differs from $z^*_{LP_k}$ by objective coefficients associated with style $\hat{i}$. The largest this gap could be is simply the difference between the maximum and minimum coefficients of variables associated with $\hat{i}$,

$$\max_{j \in \mathcal{M}}\{p_j \tilde{D}_{\hat{i},j,k}\} - \min_{j \in \mathcal{M}}\{p_j \tilde{D}_{\hat{i},j,k}\} \ .$$

Since $\hat{i}$ could represent any of the styles, the overall largest difference between $z^*_{LP_k}$ and $z^*_{IP_k}$ is

$$\max_{i \in \mathcal{N}} \left\{ \max_{j \in \mathcal{M}}\{p_j \tilde{D}_{i,j,k}\} - \min_{j \in \mathcal{M}}\{p_j \tilde{D}_{i,j,k}\} \right\} \ .$$

Therefore, we have

$$z^*_{LP_k} - z^*_{IP_k} \leq \max_{i \in \mathcal{N}} \left\{ \max_{j \in \mathcal{M}}\{p_j \tilde{D}_{i,j,k}\} - \min_{j \in \mathcal{M}}\{p_j \tilde{D}_{i,j,k}\} \right\} \ .$$

Q.E.D.

Proof of Theorem 2

We prove the first statement - "LP Bound Algorithm terminates with $z^*_{IP} = z^*_{IP_{\hat{k}}}$" - by contradiction. By construction of Step 4c, the algorithm clearly terminates, so we just need to show that when it terminates, $z^*_{IP} = z^*_{IP_{\hat{k}}}$. Assume the contrary and consider the following cases:

1. $z^*_{IP} < z^*_{IP_{\hat{k}}}$

Let $x^*_{i,j}$ be an optimal solution to $(IP_{\hat{k}})$. Set $y^*_{i,j,\hat{k}} = 1$ if $x^*_{i,j} = 1$ and 0 otherwise $\forall\ i \in \mathcal{N}, j \in \mathcal{M}$. Set $y^*_{i,j,k} = 0\ \forall\ i \in \mathcal{N}, j \in \mathcal{M}, k \in \mathcal{K} \backslash \hat{k}$. We have constructed $y^*_{i,j,k}\ \forall\ i \in \mathcal{N}, j \in \mathcal{M}, k \in \mathcal{K}$ to be a feasible solution to (IP) with the same objective value as $(IP_{\hat{k}})$. Thus, $z^*_{IP} \geq z^*_{IP_{\hat{k}}}$, and we have a contradiction.

2. $z^*_{IP} > z^*_{IP_{\hat{k}}}$

In order for this to happen, there must exist some $k' \neq \hat{k}$ such that $z^*_{IP_{k'}} > z^*_{IP_{\hat{k}}}$. Let l be the value of the index in Step 4c upon termination of the algorithm. Let $\hat{l}$ represent the index of the final value of $\hat{k}$ in the ordered set from Step 2 upon termination of the algorithm.

(a) $k' = k_{l'}$ for some $l' > l$

$z^*_{IP_{\hat{k}}} \geq LB \geq z^*_{LP_{k_{l'}}} \geq z^*_{IP_{k_{l'}}}$ $\forall\ l' > l$ which is a contradiction.

(b) $k' = k_{l'}$ for some $\hat{l} < l' \leq l$

At the beginning of the l' iteration of the loop in Step 4, we have $z^*_{IP_{k_{\hat{l}}}} = LB$. Since by definition the algorithm terminates with $\hat{k} = k_{\hat{l}}$, we must have $z^*_{IP_{k_{l'}}} \leq LB = z^*_{IP_{k_{\hat{l}}}}$, which gives us a contradiction.

(c) $k' = k_{l'}$ for some $l' \leq l < \hat{l}$

In order for $\hat{l} > l$, $k_{\hat{l}} = \arg\max_{k \in \mathcal{K}}\{LB_k\}$. Similar to the previous case, at the beginning of the l' iteration of the loop in Step 4, we have $z^*_{IP_{k_{\hat{l}}}} \geq LB$. Since by definition the algorithm terminates with $\hat{k} = k_{\hat{l}}$, we must have $z^*_{IP_{k_{l'}}} \leq LB \leq z^*_{IP_{k_{\hat{l}}}}$, which gives us a contradiction.

(d) $k' = k_{l'}$ for some $l' < \hat{l} \leq l$

Since by definition the algorithm terminates with $\hat{k} = k_{\hat{l}}$, we know that in the $\hat{l}$ iteration of the loop in Step 4, $z^*_{IP_{k_{\hat{l}}}} > LB$, and since $l' < \hat{l}$, we must have that $LB \geq z^*_{IP_{k_{l'}}}$. Thus we have $z^*_{IP_{k_{\hat{l}}}} > LB \geq z^*_{IP_{k_{l'}}}$, which gives us a contradiction.

Next we prove the second statement - "An optimal solution to (IP) is the optimal price assignment given by the solution to $(IP_{\hat{k}})$" - by constructing a feasible solution to (IP) with an objective value of $z^*_{IP_{\hat{k}}}$. Since we showed in the first part of the proof that $z^*_{IP} = z^*_{IP_{\hat{k}}}$, we know that such a feasible solution must be an optimal solution to (IP).

Let $x^*_{i,j}$ be an optimal solution to $(IP_{\hat{k}})$. Set $y^*_{i,j,\hat{k}} = 1$ if $x^*_{i,j} = 1$ and 0 otherwise $\forall\ i \in \mathcal{N}, j \in \mathcal{M}$. Set $y^*_{i,j,k} = 0$ $\forall\ i \in \mathcal{N}, j \in \mathcal{M}, k \in \mathcal{K} \backslash \hat{k}$. We have constructed $y^*_{i,j,k}$ $\forall\ i \in \mathcal{N}, j \in \mathcal{M}, k \in \mathcal{K}$ to be a feasible solution to (IP) with the same objective value as $(IP_{\hat{k}})$.

Q.E.D.