An Analytical Approach for Understanding Promotion Effects on Demand and Improving Profits

Chaitnay Singh, Xueying(Alicia) Yang, Uma Aiyer, Matthew A. Lanham

Purdue University Krannert School of Management

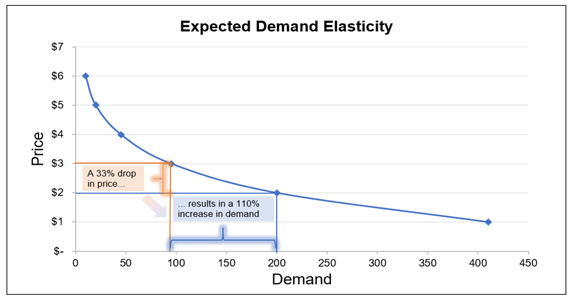
singh488@purdue.edu; yang551@purdue.edu; uaiyer@purdue.edu; lanhamm@purdue.edu

Abstract

The objective of this study is to design and develop a better revenue management system that focuses on leveraging an understanding of price elasticity and promotional effects to predict demand for grocery items. This study is important because the use of sales promotions in grocery retailing has intensified over the last decade where competition between retailers has increased. Category managers constantly face the challenge of maximizing sales and profits for each category. Price elasticities of demand play a major role in the selection of products for promotions, and are a major lever retailers will use to push not only the products on sale, but other products as well. We model price sensitivity and develop highly accurate predictive demand models based on the product, discount, and other promotional attributes, using machine learning approaches, and compare performance of those models against time-series forecasts.

Introduction

The purpose of this study is to understand the impact of historical promotions on the demand of a product and create predictive models that best predict demand based on parameters like price, discount, and holidays. Our target decision makers are Category Managers, who constantly face the challenge of determining the best price for their products that will lead to a high demand, and in turn drive up profits. The problem lies with trying to understand the relationship of the price for a particular product and how that will translate to sales.



We want to understand what is the best machine learning model that one could achieve to predict sales for grocery items that have been on promotion. Next, we test whether creating machine learning elasticity curves obtained from demand model could help us identify promotional prices better than using traditional log-log model.

Methodology

Data Sources

The data is obtained from a regional supercenter chain in the United States. The data was first split by Product ID and then grouped by Store ID within each subset. The data set consisted of roughly 1200 products from 250 stores throughout the United States, each observation contains 172 columns.

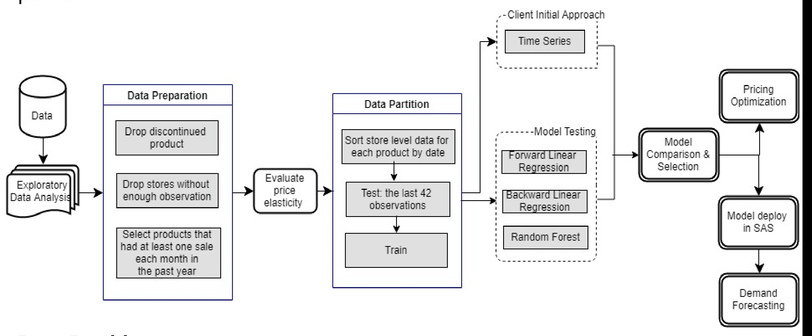
Exploratory Data Analysis

We started the initial data analysis by plotting discount rate and sales units overtime to see if the price elasticity relationship and the seasonality of sales were present as expected. In the process, we discovered that a decent portion of the products were not sensitive to price change at certain stores, and a selection of products only sell at a certain period of the year. Another important discovery is that there are certain products that have no sales for a continuous period of time. These items could be seasonal products such as Christmas decorations, or turkeys. However, it is also possible the store simply stopped carrying the product. In both situations, the data will be extremely skewed and may not be suitable for model building and prediction and was beyond the scope of this study. A basic ln(units) ~ ln(discount) regression performed to obtain price elasticity for each product at the store level.

Data Preparation

Some data transformation was done before we obtained the data. This included log transformation on the discount price and sales units, and forcing negative sale units (possible product returns) to zero. Some of the independent variables, such as “Day of Year”, “Month of Year”, ”Day of Week” were dummy transformed for model building. We also checked for multicollinearity among all independent variables and removed colinear features, such as price and sales revenue

The products and stores that did not have continuous sales in each month were eliminated from the dataset due to the lack of observations and variances in price points.



Data Partition

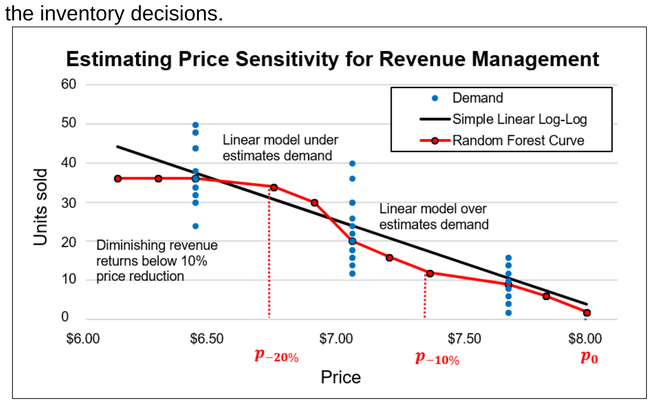
The data was divided into two different groups, the test set that comprised of 42 observations, which translates into 6 weeks of daily data, and the train set which contains the rest of the observations from the corresponding store. The size of the train set varies from 42 to 1300 observations. The train set was used to build the models. The test set was used to assess, fine-tune, and compare the models.

Model Building and Comparison/Selection

To estimate the product demand, predictive models were built on the training set using machine learning techniques, including time series models, forward selection linear regression, backward selection linear regression and random forest. These models were compared and selected for prediction on the basis of Root-Mean-Square Error (RMSE) on the test set.

Predictive Model

The corporate partner’s need was to have the most accurate prediction and interpretation was not a major concern. We decided to select the best performing model based on RMSE at the store level, because each product performs differently at each store due to the differences in customer base and geographic location. The final model was selected for each product at the store level. The best performing models were used to (1) identify the optimal promotion rate given constraints of profit margin, and (2) predict demand given a promotion strategy to guide the inventory decisions.



Model Formulation

Response variable

y_j = # of units sold $j;\ i = 1,..,N$

Predictor variables

x_j = discount price $j;\ i = 1,..,N$

w_j = week of the year $j;\ i = 1,..,N;\ w_j \in \{0,1\}$

d_j = day of the week $j;\ i = 1,..,N;\ d_j \in \{0,1\}$

p_j = promotion flags $j;\ i = 1,..,N;\ b_j \in \{0,1\}$

t_j = mean temperature $j;\ i = 1,..,N$

h_j and o_j =holiday and special occasion flags $j;\ i = 1,..,N;\ h_j \in \{0,1\}$ and $o_j \in \{0,1\}$

Linear Regression Model:

$$y_j = [\beta_{1j}x_j + \beta_{2j}w_j + \beta_{3j}d_j + \beta_{4j}p_j + \beta_{5j}t_j + \beta_{6j}h_j + \beta_{7j}o_j]$$

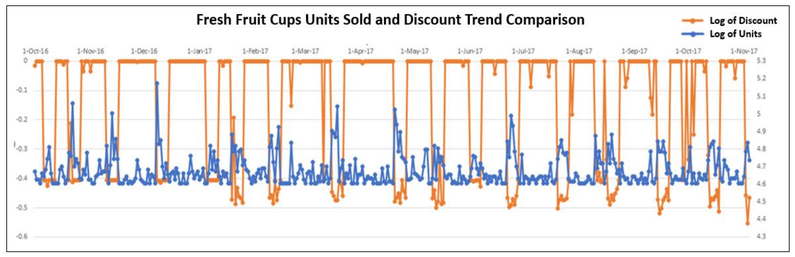
Random Forest Regression Model:

The Random Forest Regression model is an ensemble method of modeling that can be used for both classification and regression type problems. They operate by constructing a myriad of decision trees and outputting a mean prediction (in case of regression type problems) of the individual trees.

Results

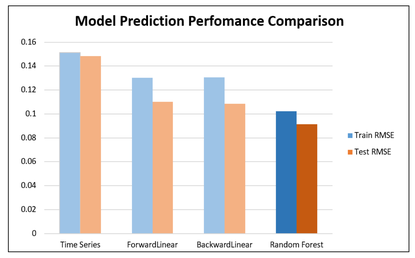
Time Series Trend

Seasonality and price sensitivity shown by discount rate and sales unit over time.



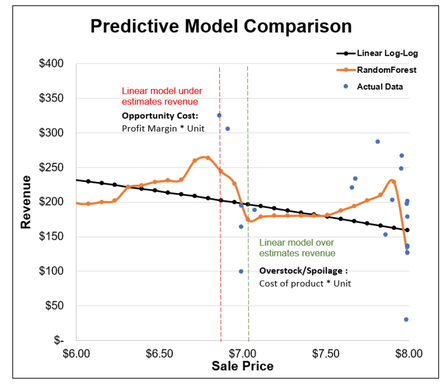
Model Comparison

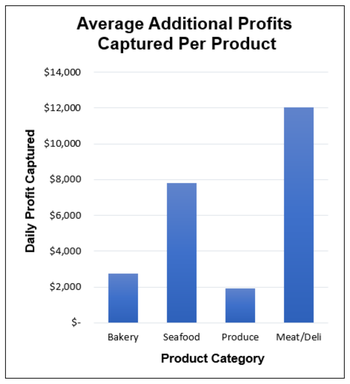
We compared time series models, two linear models and random forest. Random forests is the best model for majority of the products at each store. Linear models are used for a sizable portion of the stores and products. Overall, all three models showed significant performance improvements from the current approach in place -- time series model.



Profitability Impact Assessment

Simulations with the random forest model against linear model suggests the ability to capture demand more accurately and reduce both stock out and spoilage costs. An estimation of additional operational profits captured by using random forest model was calculated by identifying the profit loss due to understock and additional cost due to overstock and spoilage for each product across stock.





Conclusions

Evaluating the effectiveness of a promotion strategy and accurately understanding demand is a crucial component for grocery businesses to continue to drive sales while retaining acceptable margin. We were able to build fairly accurate demand predictions and identify price sensitive products. We are continuing this project to create a prescriptive tool that category managers can use to find the optimal price for a product based on the estimated demand at various pricing points. Using a traditional log-log model provides some insight about elasticity, but the relationship among demand and price is often non-linear and we posit our demand model can provide a dynamic decreasing elasticity curve that would generate more revenue and yield better profit than traditional approaches.

Acknowledgements

We would like to thank the corporate partner and our professor for their guidance and feedback throughout this project. The Purdue BIAC partially funded this work.