# Estimating Product Substitution for Grocery Items

**Michael Roggenburg, Juan Pablo Bustamante Páez, Guanzhu Mou, Matthew A. Lanham**

Purdue University Krannert School of Management

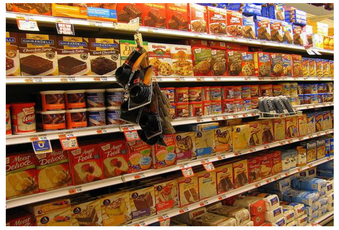
Mroggenb@purdue.edu; bustamaj@purdue.edu; gmou@purdue.edu; lanhamm@purdue.edu

## Abstract

This study investigates multi-classification predictive models to estimate product substitution among a set of grocery items in order to provide a better breadth versus depth category strategy. The motivation for this investigation is to help categorical managers better understand their customers purchasing behavior and preferences, so they can make better assortment decisions for their categories. Category management requires having a strategy that balances breadth and depth of products offered while satisfying financial and space constraints. On one end of the spectrum, one could provide extensive breadth and have many different products with few substitutes, while on the hand, one could offer a few products but have a wide depth of substitutes (e.g. different colors, different brands, etc.). In collaboration with a national grocery retailer, we build multi-classification predictive models on one category of products. These products were sold at various stores and are similar in that they are the same food item, but are different based on product attributes and brands. We investigated Multinomial Logistic Regression, Naïve Bayes, Deep Learning and Gradient Boosted Machine models and show how predictions from these models can be used to devise a category strategy to maximize sales.

## Introduction

Grocery shopping is a regular daily routine for everyone and customers are continually sensitive to substitute products within the same category. With an ever-increasing selection of brands to choose from it is important to understand what products are best to stock, with regards to the breadth and depth of each category. The trade-off between the breath and depth within a category is crucial for optimizing shelf space and sales. Therefore, it is important to understand customer purchasing behavior and predict the products chosen within each category. The reasoning for creating a model for grocery substitution is to help optimize store stocking procedures and generate a greater profit margin.



- What are probabilities for different product substitutes chosen by customer?
- What is the ideal distribution of substitutes products, breadth and depth, to maximize sales

A possible approach to solving this problem is to use a set of classification algorithms and identify product groups with the purpose of finding substitute products within each class.

## Literature Review

How can you determine which product in a grocery store should be replaced? The basic understanding of product replacement comes from data clustering.“A Logit Model of Brand Choice Calibrated on Scanner Data” a research done by two professors from the Massachusetts Institute of Technology, Peter M. Guandagni and John Little. The study determined that several variables such as brand loyalty, brand size, absence of price promotion, regular shelf price and promotional cuts were statistically significant in forecasting purchasing behavior. This research shows how certain products can be easily replaced by analyzing customers preference. However, our group try to focus more in, how did they manage to do this? They place a constant set of survey panels to administer a test group to build brand and purchasing preferences. From this a multinomial logistic regression model was fit to a set of 100 households over a 32-week period which was then benchmarked against a prediction period of 20 additional weeks.

| | Deep Learning | Multinomial Logistic Regression | Naïve Bayes | Gradient Boosted Machine |
|---|---|---|---|---|
| Our Study | ✖ | ✖ | ✖ | ✖ |
| MIT Study | | ✖ | | |

## Methodology

Figure 2 outlines the workflow for our analysis through data merging, EDA, cleaning, feature selection and removal, pre-processing, partitioning, training, validation and results.

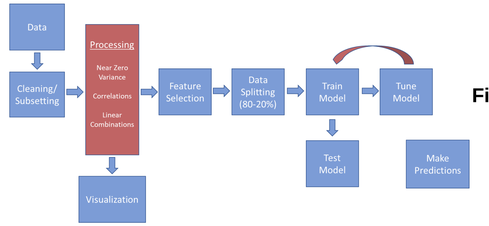


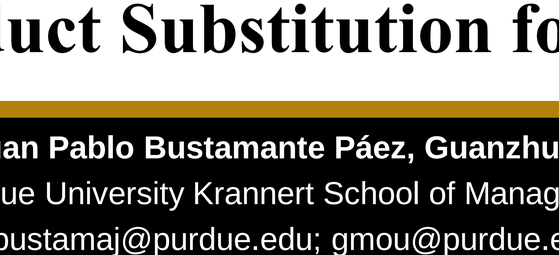

**Figure 2. Study Design**

### Data

Out data consisted of transaction history for a major grocery chain, split into 3 tables containing store location, product category hierarchies and transaction information for each product.

### Data Cleaning & Pre-Processing & Feature Selection

We first merged out data by product category, brownies, and by a particular store location and began with 181 variables. Next we separated the data into 3 tables each with separate timelines relating to the changes in the product category and removed unimportant columns. Near zero variance categorical variables were removed, linear combinations and correlated numeric features were removed and range standardization, Yeo-Johnson transformation and PCA were performed. Finally, our Team removed several categorical variables related to the date of purchase and managed to reduce the data to 10 key variable, with each dataset containing roughly 500 observations.




### Methodology

Our team chose to use the H2O package in R to perform analysis on our data. Using a 80-20% split for the train/test set and 10-fold cross validation to account for small datasets. The four models that we chose to test were a Multinomial Logistic Regression, Naïve Bayes, Deep Learning and Gradient Boosted Machine. Each model is capable of approaching a multiclass problem and has its own advantages.

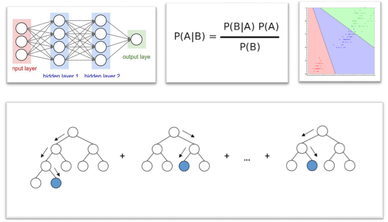



**Figure 3. Deep Learning Model (Top Left), Naïve Bayes Classifier (Top Middle), Multinomial Logistic Regression (Top Right), Gradient Boosted Machine (Bottom)**

Multinomial models generalize logistic regression to multiple factors and is a good overall model for this problem. Naïve Bayes models use assumptions form prior distributions are generally able to scale with most problems. The Deep Learning model provides the most flexibility with the advantage of adding multiple hidden layers, altering node amounts and its inherent robustness in dealing with noisy data. Finally, the Gradient Boosted Machine is a great general forward learning ensemble method and, like the Deep Learning model it is very robust to random features.

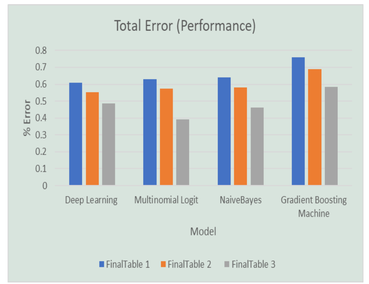
### Model Evaluation / Statistical & Business Performance Measures

All four predictive models were measured on their error rate for classification. For our business problem, it is essential to correctly predict which product a consumer will purchase at the shelf, so a low error rate will provide the best picture of what brands and products consumers want.

## Results

Figure 4 and 5 represent the total error found in all 4 models. The table with light green background analyzes the error from the performance, the table with light yellow background represents the error of the train model. Every color column embodies each individual FinalTable. According to this data FinalTable3 has the most optimal results.

**Figure 4. & 5. Model Evaluation**



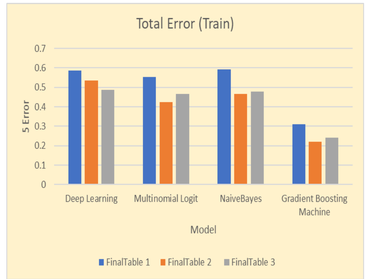
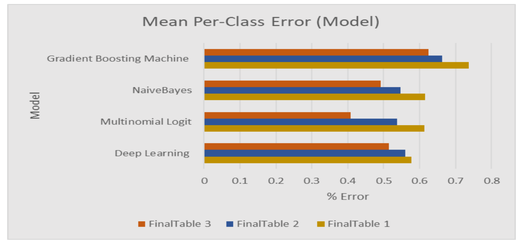




Figure 6 represent the Mean Per-Class Error of each Model. According to the results the FinalTable 3 found the most optimal substitutes for each product studied. However, the results are not as strong as expected, the error rates from certain models and tables are too high to drive any conclusions from them.

**Figure 6. Mean Per-Class Error (Model)**




## Conclusions

Understanding product substitution continues to be a pressing issue for store stocking managers. Real insight into the ideal balance of breadth and depth for a product category can help optimize the products stocked in that group.

Through our research of grocery transaction data, our group believes that a model can be made to predict optimal stocking for substitute products given an ideal product category. Given that our team analyzed brownie sales specifically, we believe that a different category such as condiments with more breadth and less depth. However, through our research we were able to develop a set skeleton of models for dealing with several types of product categories. These models can be used to generate predictions for a given product, within a category, in a transaction relating to that category. This can help category managers optimize shelf capacity by stocking more of a brand that has a higher probability of being purchased and give insight on the ideal balance of breath and depth.

## Acknowledgements

We thank Professor Matthew Lanham for constant guidance on this project.