# Predicting Brand Loyalty in Grocery Shoppers

Rafael Rivera-Soto
rivera43@stanford.edu

Daniel Gardner
dangard@stanford.edu

*Abstract*—**We predict a household's propensity to purchase national brand name products based on a number of demographic factors, including age, education, race, and income. We find that higher income, older, and larger households are more likely to purchase brand name products as opposed to generic store brands. We use households and purchases from the Nielsen scanner dataset, and employ various machine learning algorithms to make predictions. Using a categorical variable to represent a household's brand loyalty, we are able to achieve a 96% prediction accuracy on our test set of households across 25 products.**

## I. Introduction

Brand loyalty is a chief concern for marketers of grocery products. Consumers will often buy the same brand of a household good for their entire lives. It is vital for product marketers to determine which consumers are especially 'brand loyal' so that the marketers can target advertising and promotions toward them. For most grocery items, consumers are presented with the choice of purchasing a brand name product or a store brand product. Well-known national brands include Kellogg's cereal, Heinz ketchup, and Tide detergent, while store brands include Costco's Kirkland, Target's Market Pantry, and Walmart's Great Value. Despite a brand product often being more expensive than its store brand counterpart, many consumers prefer the reliability and known quality of the national brand. This brand preference or loyalty widely varies across different households and especially across different product types. We use machine learning techniques to predict a household's preference for brand name products and determine what factors into that predilection.

For a given product, we label households as 'brand loyal' or not based on the brand ratio of their past purchases. Our algorithm takes as inputs the demographics of a household along with product-specific parameters. We then use logistic regression, support vector machines, and adaptive boosting to predict the household's brand loyalty for that product. We also experiment with using the k-nearest neighbors algorithm to find similar product clusters and utilize these clusters as a feature in our final brand loyalty prediction.

## II. Related Work

Brand loyalty has been extensively studied by economists and marketing researchers [1] [2]. Often their research focuses on a specific demographic group and observes whether this group has different behavior than the population at large. Bronnenberg et al. [3] examines grocery purchases by consumers who are particularly well-informed about the homogeneity of certain brand and store-brand products and observes that they purchase the cheaper store brand more often than the average shopper. They are able to do this by matching employment information (choosing medical professionals and chefs) with domain-specific products (pain medication and baking goods). In a more recent paper, Bronnenberg et al. [4] examines households that have lived in multiple regions of the United States in their lifetime and finds that the 'brand capital' they have developed in the past makes them have different brand loyalties than similar consumers in their current region.

The most common use of machine learning algorithms in consumer behavior research is to create market baskets', or products that are frequently purchased together [5]. This is useful when trying to develop marketing campaigns to mesh multiple product categories, but it does little to explain which consumers might prefer to buy a brand of a particular product. As evidence of the increasing interest in applying machine learning to the field, the popular data science website kaggle.com has hosted multiple competitions to develop models related to grocery purchases [6] [7]. These include a problem posed by a marketing research firm to predict when shoppers will visit a store next and how much they will spend, as well as a problem from Walmart to classify different types of shopping trips.

The previous approaches to brand loyalty have studied specific groups and how they behave differently, while machine learning in the grocery space has dealt mostly with clustering of substitute and complementary products. We will instead focus on what characteristics of the average consumer contributes to his or her brand purchasing choices. This analysis allows us to address the most pressing question for marketers - which slice of the population they should focus their limited advertising revenue on to maximize the success of their brand, both in the short term and long term [8].

## III. Dataset and Features

### A. Dataset

We use the Nielsen Consumer Panel Dataset from the James M. Kilts Center for Marketing at the University of Chicago Booth School of Business. The Nielsen Company provides scanners to households who keep track of their purchases at grocery stores. This data contains more than three million unique universal product codes (UPCs) from transactions between 2004 and 2014. Households have unique ID numbers and usually remain in the dataset over multiple years. The approximately 60,000 households are located in 50 different major metropolitan areas in the United States, representing a broad spectrum of consumers across the country. For this project we used only the 2014 data, as it contains the most unique UPCs.

The data is stored in four separate datasets: Panelists, which contains all the demographic information about each household; Trips, which contains all shopping trips for all the households; Purchases, which contains all the products purchased and the price paid on all shopping trips; and Products, which connects each UPC with one of 1400 product modules.

Each product module represents a group of UPCs that are essentially substitutes for one another. Examples include CEREAL - READY TO EAT, SEAFOOD-TUNA-SHELF STABLE, DAIRY-MILK-REFRIGERATED, and PAIN REMEDIES - HEADACHE.

### B. Preprocessing

In order to prepare our data to use in our machine learning algorithms, we have to determine the purchase history of each household for a given product, and also calculate certain metrics for that product. An example is helpful for illustration:

To get household purchase history for product module 5000, we filter the Products dataset for UPCs with product module = 5000. We then merge that UPC file with the Purchases dataset to get a list of all purchases made from product module 5000. We then merge with Trips and then with Panelists and now have a list of all purchases, each labeled with its purchasing household. We then sum the number of purchases each household makes in total and also for the specific brand label = 'CTL,' which is the brand code for store brand. We can calculate the number of brand name products they bought by taking $\#total\ purchases - \#CTL\ purchases$. The household's 'brand loyalty' is simply $\frac{\#brand\ purchases}{\#total\ purchases}$.

We then calculate the following metrics for each product:

$$\begin{aligned} average\ unit\ price &= \frac{total\ \$\ spent}{\#units\ purchased} \\ brand\ price\ ratio &= \frac{average\ brand\ unit\ price}{average\ store\ brand\ unit\ price} \\ brand\ purchase\ ratio &= \frac{\#\ brand\ purchases}{\#\ total\ purchases} \end{aligned}$$

We choose a collection of 25 representative products to make predictions about. We require each product to have more than 10,000 purchases in the year and we include a mix of products across departments, from FROZEN FOODS to ALCOHOLIC BEVERAGES to HEALTH & BEAUTY. We are especially careful to select products that have a variety of household brand loyalty ratios. In general, most households were almost 100% 'brand loyal' or not for any given product. It is important to select products that at least have some households that are 100% 'brand loyal' and 0% 'brand loyal' so that our binary prediction models are meaningful. Some of our selected products are strongly 'brand' (detergent), some have mostly off-brand purchases (milk), while others are a mix of both (eggs) (Fig. 1 & 2).

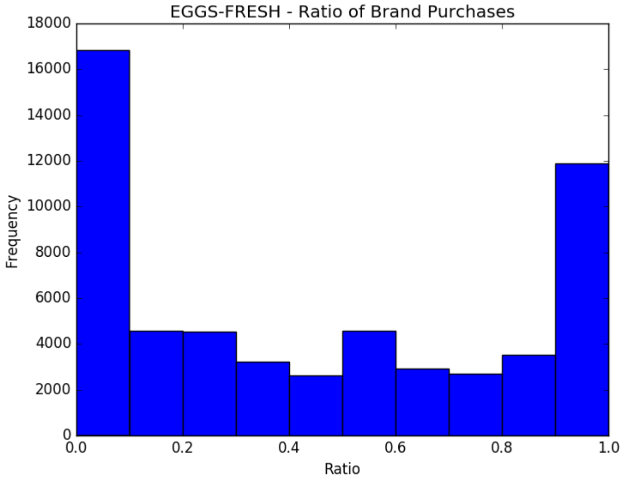

Fig. 1. Brand vs Non-Brand Ratio for Eggs

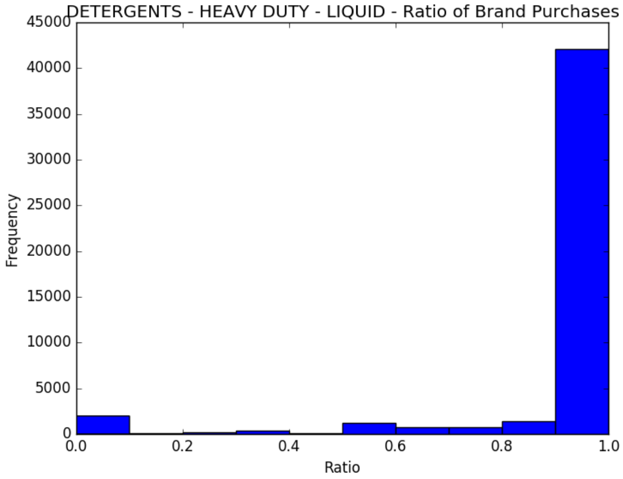

Fig. 2. Brand vs Non-Brand Ratio for Detergents.

### C. Feature Selection

The demographic information about each household in the Panelists file is quite extensive. It contains the number and age of adults and children, the education and employment information of the male and female head of house, the zip code and residence type, and the household's race and income. It also contains information about the presence to kitchen appliances, televisions, and internet connection in the home.

For our machine learning algorithms, we chose as features income, race, household size, age and presence of children, and head male and female education and employment status. Each of these was a categorical variable and required us to create a dummy binary variable for each of its categories. For example, age_and_presence_of_children has eight different categories representing combinations of numbers and ages of children in the home. We reduced this to three binary categorical variables: has_young_children, has_children, has_teenagers. Age and income did not require dummy vari-

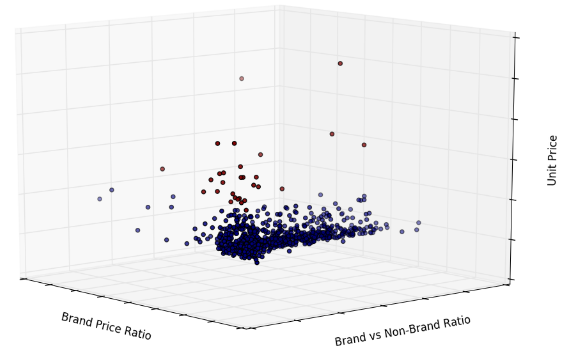

Fig. 3. Two product clusters extracted using the KNN algorithm. Unit price heavily influences what cluster products fall into.

ables because their categories were granular and monotonically increasing. One notable feature we omitted was occupation, which simply had too many categories and mostly correlated with income.

## IV. Methods

### A. K-Nearest Neighbors for Product Clustering

We use the K-Nearest Neighbors (KNN) algorithm to cluster our selected products and other products in the same category. The KNN algorithm takes as input a training dataset and groups the data into "clusters". In our implementation, we extracted three product features to base our clustering on: unit price, brand price ratio and brand purchase ratio (Fig. 3). These clusters are then used as one of the features for our Logistic Regression model. The idea of this feature is that a person's buying behavior might be influenced by what kind of products they're buying.

### B. Logistic Regression

Our initial efforts were concentrated upon whether consumers tend to buy more branded or non-branded products. This is a binary classification task for which we implemented a Logistic Regression model. A Logistic Regression squashes the output of the model in the range $y = \{0, 1\}$ using the sigmoid function (Eq. 1). The output values are then interpreted as the probabilities, thus any output greater than 0.5 is classified as belonging to the positive class and to the negative class otherwise (Eq. 2).

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}} \tag{1}$$

$$\begin{aligned} P(y = 1|x; \theta) &= h_\theta(x) \\ P(y = 0|x; \theta) &= 1 - h_\theta(x) \end{aligned} \tag{2}$$

We based our initial predictions on products whose distributions between branded buyers and non-branded buyers was almost even. We labeled someone as being on the "branded-buyer" class if they're above the ratio cut-off value of 0.5, all other consumers were labeled as belonging to the"non-branded" class.

### C. Support Vector Machines

The Support Vector Machine is a discriminative classifier that finds an optimal hyperplane with the largest margin between the classes. These models also allow us to implicitly map our input features into a high-dimensional feature space where the optimal hyper-plane might result in a better division between the classes. These feature mappings are called kernels. In our implementation, we used the Radial Basis Function (RBF) kernel (Eq. 3). SVMs use a particular choice of the loss function called the "Hinge Loss" (Eq. 4). To fit this model, we use an algorithm such as Stochastic Gradient Descent to adjusts our weights in such a way that the Hinge Loss is minimized. We used this model in the binary classification model described above and compare the effectiveness.

$$K(x, x') = exp(-\frac{||x - x'||^2}{2\sigma^2}) \tag{3}$$

$$\delta(z, y) = max\{0, 1 - yz\} \tag{4}$$

### D. Boosting

The idea of this model is to take a weak learning algorithm, that is, any learning algorithm that does slightly better than random and transform it to a strong classifier that does much better than random. Roughly, this method begins by assigning every training example equal weight. It then receives a weak-hypothesis that does well according to the current weights. A weak hypothesis is an algorithm that takes as inputs some distribution (weights) $p$ and outputs a weak learner that does better than random (Eq. 5). After evaluating the results after incorporating the new hypothesis, it re-weights the examples in such a way that incorrect classifications receive higher weights and correct classifications receives lower weights. In this way, boosting is able to create a strong hypothesis that generalizes well to new examples.

$$\sum_{i=1}^{m} p^{(i)} 1\{y^{(i)} \neq \phi_j(x^{(i)})\} \leq \frac{1}{2} - \gamma \tag{5}$$

### E. Multinomial Logistic Regression

After our initial efforts, we decided to extend our model to incorporate more than one class. In particular, we chose to divide consumers into three bins using cutoffs at 0.33 and at 0.66. This resulted in the following class division:

$$C = \begin{cases} 0, \text{if } ratio < 0.33 \\ 1, \text{if } 0.33 \leq ratio < 0.66 \\ 2, \text{if } ratio \geq 0.66 \end{cases} \tag{6}$$

All of our previous models have been binary classifiers. In order to account for more classes we fit a Softmax Regression model. This is a generalization of the Logistic Regression models to multiple classes. In particular, Softmax Regression

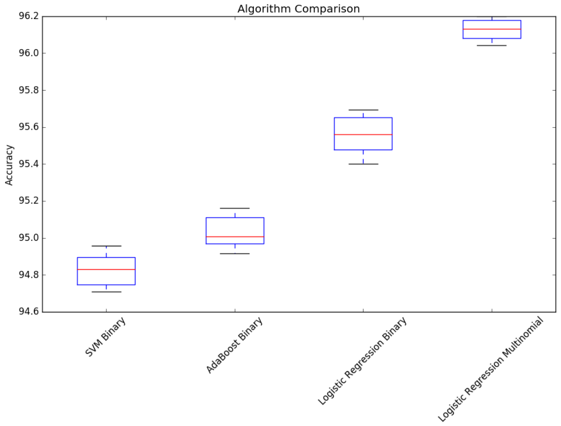

Fig. 4. Comparison of Algorithms used for Classification. Multinomial Logistic Regression yielded the best results.

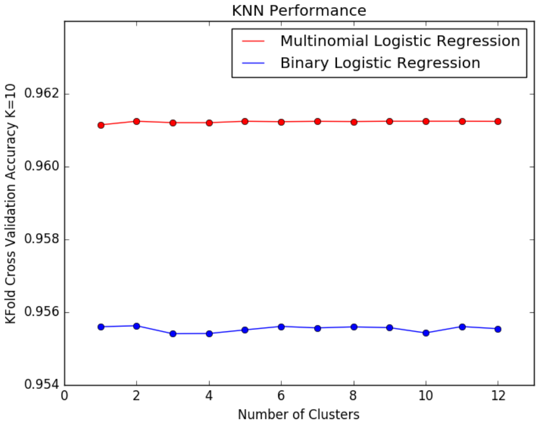

Fig. 5. Analysis of the effectiveness of the Cluster feature. K=2 yielded the best results on both the binary and multinomial case.

uses the Multinomial Distribution. The probability that our features take on a certain class is given by Eq. 7.

$$p(y = i|x; \theta) = \frac{e^{\eta_i}}{\sum_{j=1}^{k} e^{\theta_j^T x}} \tag{7}$$

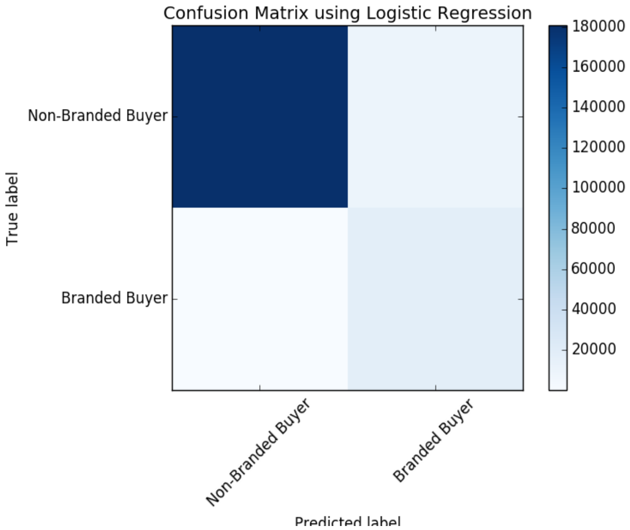

Fig. 6. Confusion Matrix for Binary Prediction

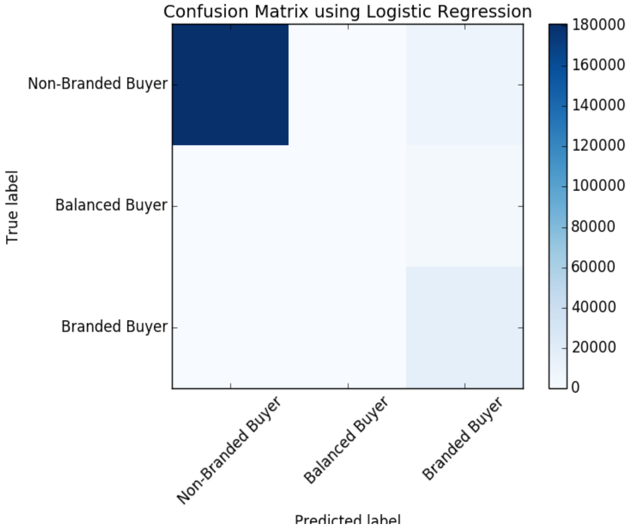

Fig. 7. Confusion Matrix for Multinomial Prediction

| Overall | Eggs | Bacon | Contraceptives | Ketchup |
|---|---|---|---|---|
| Income | White | Income | Income | Income |
| Household Size | Female Education | Female Age | Household Size | White |
| Female Age | Income | Male Age | Age/Presence of Children | Female Age |

Fig. 8. Most effective features derived from the recursive feature elimination method.

## V. Experiments

Using the methods described above, we ran several experiments for both the binary and multinomial case. Our initial experiments were concerned with getting baseline results for Logistic Regression, SVMs and Boosting on the binary classification case. Once this was done, we chose the best model of the three and experimented on the multinomial case. All models were evaluated using K-Fold Cross Validation with 10-folds. K-Fold Cross validation is a method for assessing how results will generalize beyond the training set. It partitions the dataset into K equal sized folds. Of the K folds, one is retained for testing and K-1 are used for training. This process is repeated K times with each of the folds being used once for testing. The results are then averaged to produce a single estimation. For our dataset, this results in about 630,000 examples used for training and 70,000 used for testing.

We also performed Recursive Feature Elimination (RFE) to evaluate which features were most impactful toward our classification. RFE is initially trained on every feature and assigns weights to each of the features. Then, features whose weights are the smallest are pruned from the current set of features. This process is then repeated recursively.

### A. *Results*

Our results, for the binary case show Logistic Regression to be the best performing algorithm on the binary case (Fig. 4). It yielded an accuracy of 95.5% on average over all folds while SVM and Boosting yielded an accuracy of 94.8% and 95% respectively. We then chose Logistic Regression to predict on the multinomial case, this yielded an accuracy of 96.1%.

We performed an analysis of the effectiveness of our clustering feature on both the binary and multinomial case for Logistic Regression. The results can be seen on Fig. 5, they show that we get the best result when $K = 2$.

### B. *Discussion*

Multinomial logistic regression has the best accuracy, which makes sense because it allows for a more robust representation of households' brand loyalty. The regular logistic model performs slightly worse because it only accounts for households who are on either extreme of the spectrum, while the multinomial establishes a class for the brand neutral consumer who buys a good mix of national brand and store brand products. We considered extending this to four or five categories, but decided that three is most logical for labelling a household. The logistic confusion matrices show that our selected products ended up having much more store brand purchases than we originally intended, but our high accuracy is still reflected in individual products that have a good distribution (Fig. 6 & 7).

Income, Head Female Age, and Household Size were most predictive of brand loyalty (Fig. 8), which is somewhat logical. Brands cost more, so higher earners can afford to pay a premium while lower earners opt for cheaper store brand replacements. Older females often make most of a household's grocery purchases, and they are likely to have settled on a reliable brand. Household size should be explored further, as it is unclear why a larger household indicates greater brand loyalty. It could be that this correlates with larger families, who rely on making quick shopping trips and do not want to 'taste test' unknown store brands.

An interesting discovery is that certain premium products (e.g. brand eggs) are preferred by households with white, educated women. Education and race were among the least predictive factors in our overall recursive factor elimination, but they jumped to the top in a few products. This is somewhat of an extension of the pattern observed in Bronnenberg et al. [3] that certain more informed consumers favor store brands when buying homogenous products. However, it could be stated that premium products actually have qualities in their brand varieties that make informed shoppers prefer them (e.g. organic or cage-free eggs).

Drawing a direct link between a head of house's age, education, or work and brand loyalty is challenging because our dataset does not indicate which member of the household made various grocery purchases. Having this information would allow for examining how a person's demographics affect their spouse's buying decisions and vice versa.

We were surprised that product cluster and our calculated product features did little to improve our overall accuracy. Although both regular and multinomial logistic regression perform best with two clusters, the accuracy difference is only about 0.1 %. When we instead use the product features directly in the models, a similar result is observed.

One potential flaw in our approach is that we fail to account for location effects on our households' brand-buying tendencies. Different parts of the country have more or less access to different brands, and some areas have access to superior store brands [2]. Also, the same income can put a household in drastically different socioeconomic levels in different metro areas, so some normalization of income may have been helpful.

## VI. Conclusion

In general, marketers of national brand products should focus on households with high income and older female heads as potential new customers. Our machine learning algorithms showed that these demographic features are most predictive of brand loyalty. However, these people may already be loyal to another brand in the same product module and marketing efforts to get them to switch will be difficult. It might be more useful to instead predict and identify potential customers of your brand.

Future machine learning work could focus on finding a marketing strategy that attracts new customers to a brand. We could predict which groups of consumers will be most likely to switch from store brand to national brand products and then calculate potential profits from these new consumer purchases.

## References

[1] W. A. Kamakura and G. J. Russell, "Measuring brand value with scanner data," *International Journal of Research in Marketing*, vol. 10, no. 1, pp. 9 – 22, 1993. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0167811693900303

[2] B. J. Bronnenberg and J.-P. H. Dubé, "The formation of consumer brand preferences," National Bureau of Economic Research, Tech. Rep., 2016.

[3] B. J. Bronnenberg, J.-P. Dubé, M. Gentzkow, and J. M. Shapiro, "Do pharmacists buy bayer? informed shoppers and the brand premium," National Bureau of Economic Research, Tech. Rep., 2014.

[4] B. J. Bronnenberg, J.-P. H. Dubé, and M. Gentzkow, "The evolution of brand preferences: Evidence from consumer migration," *The American Economic Review*, vol. 102, no. 6, pp. 2472–2508, 2012.

[5] J. McAuley, R. Pandey, and J. Leskovec, "Inferring networks of substitutable and complementary products," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 785–794.

[6] "dunnhumby's shopper challenge," https://www.kaggle.com/c/dunnhumbychallenge, accessed: 2016-12-16.

[7] "Walmart recruting: Trip type challenge," https://www.kaggle.com/c/walmart-recruiting-trip-type-classification, accessed: 2016-12-16.

[8] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.