Predicting Brand Loyalty in Grocery Shoppers

Daniel Gardner and Rafael Rivera-Soto

dangard@stanford.edu rivera43@stanford.edu

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. For a given product, we classify households as 'brand loyal' or not, and then use machine learning techniques to model demographic characteristics and product similarities that contribute to their loyalty or lack thereof. We find that household income, family size, and head female age are most predictive.

Data Preprocessing

The Nielsen Consumer Panel Dataset has more than three million unique grocery product barcodes and groups them into 1400 product modules (milk, cereal, toilet paper, etc). We use the 2014 purchases data, which contains purchases from 60,000 households.

I. Deriving Prediction Values

We determine a household's brand loyalty for a given product module using the following formula:

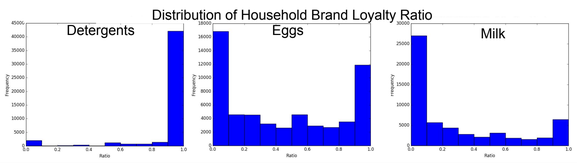
Brand Loyalty = # brand purchases / # purchases

II. Creating Product Features

In order to calculate clusters of similar products, we create three features for each product: average cost, average brand/off-brand price ratio, and brand/ off-brand total purchases ratio.

III. Selecting Representative Products

We choose 25 products to make predictions about, all of which have a large number of purchases. Some are strongly 'brand' (detergent) while others have more off-brand purchases (eggs, milk).



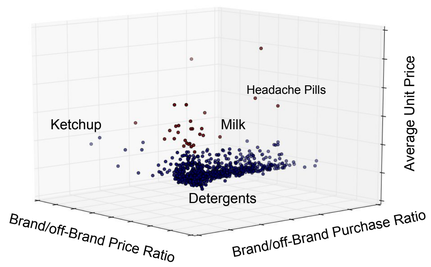
IV. Selecting Household Features

We choose 15 of the more than 40 features in the household data to use in our model. These include race, education, and income. We create dummy variables for the categorical variables.

Methods

I. KNN for Product Clustering

We use the k-nearest neighbors algorithm to cluster our selected products and other products in the same category. The clusters are used as one of the parameters for logistic regression.



II. Training Matrix

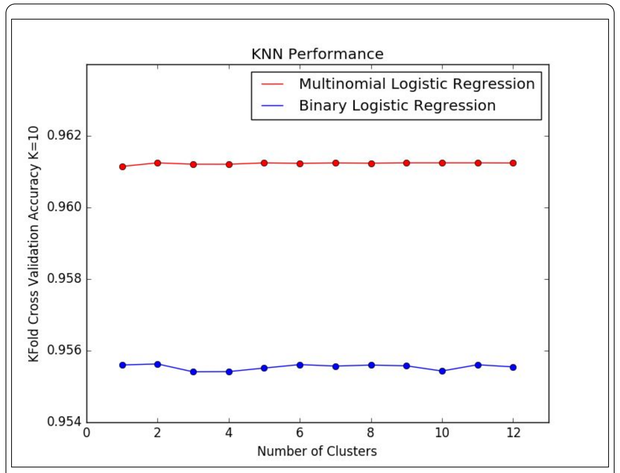
Our training data includes a row for each household-product pair. The features are the household demographics and the product cluster. We use k-fold cross validation in our testing of the models with k = 10. This results in a training set of 630,000 and test set of 70,000 for each fold.

III. Logistic Regression

We first perform simple logistic regression, treating a household's actual brand loyalty as 0 or 1 based on a .5 cutoff. We run logistic regression using k = 1-12 for the product clustering feature and find that we have highest accuracy at k = 2.

IV. Multinomial Logistic Regression

We replace the .5 cutoff with cutoffs at .33 and .66, allowing us to have a third classification for 'brand neutral'. This performs slightly better than regular logistic regression and also has its highest accuracy at k = 2 clusters.



V. Feature Evaluation

We use recursive feature elimination to evaluate the highest-weighted demographic features (product cluster was top overall). We list the common case as well as some interesting examples.

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

VI. Support Vector Machines

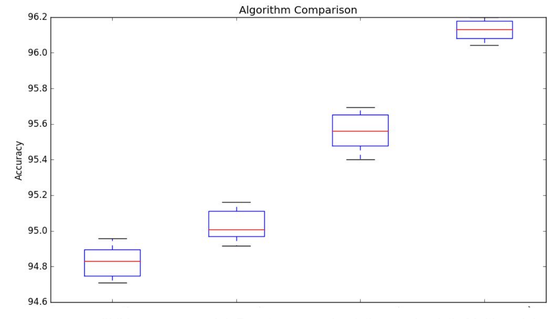
We use a Support Vector Machine classifier with the RBF kernel as one of our baseline implementations. We initialize C to be 1.0 and let gamma be 1 / number of features. This model yielded an accuracy of 94.82% on the binary classification problem.

VII. Adaptive Boosting

The Adaptive Boosting classifier uses up to 200 estimators. This classifier yielded an accuracy of 95.1% on the binary classification problem.

Results

The plot below shows our relative accuracy using cluster size k = 2. Logistic Multinomial did best.



Discussion

Multinomial logistic regression allows us to capture the middle group of brand neutral consumers, so it makes sense that this has the highest accuracy. The demographic features contributing to brand purchases are somewhat logical. Brands cost more, so higher earners can afford to pay a premium while lower earners opt for off-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 what that indicates about brand loyalty. An interesting discovery is that certain premium products (e.g. brand eggs) are preferred by households with white, educated women.

Future 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 off-brand to brand products and then calculate potential profits from these new consumer purchases.

References

Bronnenberg, et al., 2015. "Do Pharmacists Buy Bayer? Informed Shoppers and the Brand Premium." The Quarterly Journal of Economics, Oxford University Press, vol. 130(4), pages 1669-1726.