Team Name: Kittens are Cute

Random Forest Models for Predicting Egg Purchase

Phase 1: Exploratory Analysis

Upon downloading the data, we noticed that there were two distinct groups of households: households that offered demographics data, and households that did not. In order to potentially utilize demographics data for the households for which it was available, we pre-emptively split the data into two groups. "Group 1" contained households that did not offer demographics data, and "group 2" contained households that offered it. At first, the two groups appeared similar (in terms of transaction cost and savings data).

After the problem was released, we split the data by date, intending to use transactions from days 500-697 to build a mode to predict egg purchases for days 698-704, for which we had data. At this point, we noticed that group 2 made considerably more egg purchases than group 1; 61 of 629 (9.7%) of group 1 households purchased eggs during the last week, whereas 64 of 338 (18.9%) group 2 households did the same. This disparity prompted us to build different models for group 1 and group 2 households.

Phase 2: Preprocessing

For days 500-697, we calculated the avg. transactions per day, avg. quantity bought per day, avg. quantity bought per purchase, avg. spending amount, avg. spending amount per item, avg. discount, proportion of items bought from each department, avg. eggs purchased per day, days since last egg purchase, and days since last transaction for each household in each of the two groups. We also calculated whether or not the household bought eggs during the last 7 days.

Phase 3: Training

We chose to use the RandomForest R package to build our predictive models. Our number of trees was set to the default 500, and the number of predictors sampled at each node was set to 16. (Parameter tuning with the caret package in R indicated that a higher number of predictors slightly improved our model's accuracy for group 2 and offered more precise probability estimates overall.)

Results

Our two random forest models did not perform much better than the baseline in each group (compared to ZeroR, which always predicts the most common category). The model for group 1 obtained a 90.3% accuracy determined through five-fold cross validation by predicting "no" to "purchased_eggs" every time. The model for group 2, however, obtained an accuracy of 81.7%, actually correctly predicting the purchase of eggs in a few instances. Nevertheless, each model offered insight as to which factors increased the likelihood of egg purchase.

Variable Importance and Conclusions

The six most important predictors in each model are shown in the bar charts below. The most important predictor of whether or not someone will buy eggs is if they generally buy many eggs. The presence of the DELI category indicates that Kroger may be able to sell more eggs by placing eggs close to the deli section, or by marketing eggs alongside deli products.