Report for CMU Data Science Cup 2016

Team The Third Place

I. Methodology

Our data analysis is composed of the following steps:

- Cleaning and pre-processing data
- Exploratory analysis
- Model experiments
- Feature engineering
- Analysis

II. Cleaning and pre-processing data

We took multiple steps to first clean and process our data. Firstly, we noticed that there were multiple entries in which the QUANTITY, as well as other relevant variables like the BASE_SPEND_AMT and NET_SPEND_AMT were all 0. This made up approximately 3000 of the 300,000 total rows of our dataset. Furthermore, we normalized fields like the DAY from 500 to 700 to 0 to 20. This makes our data more palatable to us and the models. Finally, we added a boolean field, ”GET_EGGS”, that represented whether or not the current entry was an egg purchase or not.

Then, we took each grocery transaction, and aggregated them together into tables representing transactions by day, and then transactions by month, per household. This was done by taking a particular day/month, aggregating the QUANTITY and SPEND fields, and appending meta-data for the purchases.

III. Exploratory analysis

This is largely done prior to the release of the question, and a HTML page in the visu folder contains our work. Primarily, we tried to understand the the customer segmentation of the shop.

IV. Feature engineering

For feature engineering, we explored the relationships between different pairs of variables, in order to decide what features would be relevant to the task. Furthermore, we had to take into account the complexity of the features, because too complicated features would likely lead to overfitting, and would be hard to iterate over the short time of the contest. In the end, we decided upon looking at quantity and money spent for that particular day, whether or not the household bought eggs in the previous week, and discounts.

V. Model Experiments

In this competition, we explored the results of three strong state-of-the-art methods in classification: logistic regression, random forests, and Gradient Boosting Machines.
We trained each model with the same set of features:

VI. Baselines

We have two baseline models used for comparison. The first one is based off the observation that our output parameter of the data is heavily skewed. That is, the probability that any given week and any given household has a purchase of eggs is about 0.133. Thus, a baseline model that predicts no eggs all the time will be skewed towards a high brier score.

The second baseline is trying the probability that each household purchasing eggs as a Bernoulli random variable, so we have a binomial distribution in this case. The conditional probability is given by:

$$\begin{aligned} P\{GET\ EGGS\} &= \frac{\sum_i^{household}\sum_j^{day} GET\ EGGS_{ij}}{\sum_i^{household}\sum_j^{day} 1} \\ P\{GET\ EGGS|\ week\} &= 1 - P\{GETNOEGG\} \\ &= 1 - (1 - P\{GET\ EGGS\})^7 \\ &= 0.126284 \end{aligned}$$

VII. Experiments results

Models and Brier score	
Logit	0.14
RandomForest	0.11
XGBoost	0.09
Baseline	0.0009

VIII. Conclusion

Of the three models that we tested, XGBoost (Gradient Boosting Machines) worked the best. However, because of the skew of the data, the baseline still performed much better than all the models. Further directions for research include more feature engineering, and using a different metric like Precision/Recall for the evaluation of the model.