Tartan Data Science Cup Episode II — Polar Bears Report

Michael Rosenseberg
mmrosenb@andrew.cmu.edu
October 9, 2016

When I first started to analyze the data during the weekend, I recognized the potentially high volatility in number of baskets over time (see Figure 1). I also recognized that many of the demographic variables were left missing, which suggests that I would need to spend some time considering how to handle the large amount of missing observations in that section. At the beginning of today, I recognized that predicting next week's egg purchases would require me to consider some of the time-dependent aspects of purchasing over time in this dataset. Thus, I decided to first process the dataset into panel format (see `../codeNotebooks/getHouseholdFrame.py`).

I considered the following for my models to decide upon:

- I trained all of my models on a training set of 70% of the households and I tested all of
- I first decided to use two logistic models: one fit to the data without demographic variables and one fitted to the data with demographic variables.
- I then tried to fit one logistic model on the whole training set with the variables described in the table below.
- I chose to fit a gradient boosting machine with $eggsPurchased_{i,t}$ (binary for if household i purchased eggs on week t) on its three lags and the number of overall trips a household has made ($numBaskets_i$).
- I tried to fit a non-parametric regression with the variables suggested above, but it was too computationally expensive.

My final model was a logistic regression on $eggsPurchased_{i,t}$ for household i on week t with these coefficients:

Parameter	*Estimate*
β_0	-2.943
$\beta_{eggPurchase_{i,t-1}}$	.01410
$\beta_{eggPurchase_{i,t-2}}$	$2.046 \times (10)^{-3}$
$\beta_{eggPurchase_{i,t-3}}$	.5862
$\beta_{numBaskets_i}$	.6651

We see that all of our lags and the number of baskets (or trips to the grocery store) a household has at this grocery store increase the probability of purchasing eggs next week. This suggests that if we wanted to increase the purchasing rate, we either want to increase the number of times an individual goes to this store ($numBaskets_i$) or we wanted to optimize purchasing rates based on the monthly cycle of egg purchases.

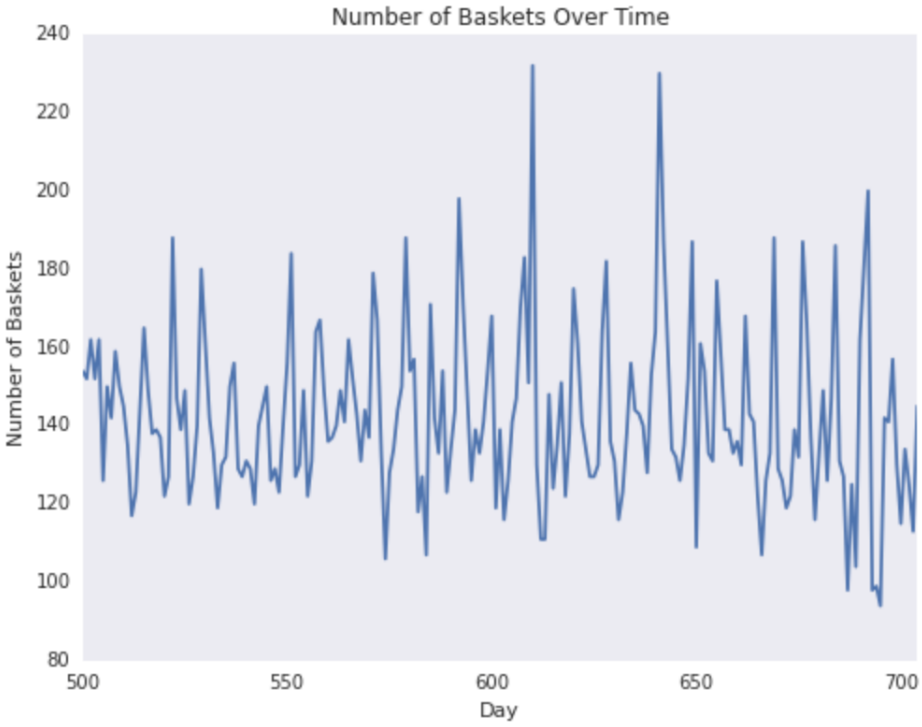

Figure 1: Number of Baskets over time.