Project One Proposal: Instacart Online Grocery Shopping

Philip Sparks (advisor: Simon Worgan)

August 8, 2017

1 Problem

The Instacart online grocery shopping dataset gives several interesting problems to solve. A primary problem will be to predict which products a customer will purchase next, correlated to their previous purchases. Secondary goals from this include what factors are occurring to retain or drive away customers. For those customers that we are retaining, what products could Instacart push to upsell customers? For those customers that are leaving, what products could Instacart sell that would bring back those previously absent customers?

2 Business Cases

Instacart is the obvious customer, as this dataset originated with them. However, lots of clients would be potentially interested in this problem. Amazon, with their recent purchase of Whole Foods, is a very interested client in this problem. Many local groceries, even if they do not have a significant online presence, should be paying attention to this data, as it contains the vast majority of their products. When one generalizes the data to how models can be build to predict and drive customer behavior, any large retailer, from Wal-Mart to your local jeweler, may find takeaways from understanding this problem.

3 Data

The Instacart dataset can be found here: Instacart Order Dataset
A helpful description of the schema is also found here: GitHub Gist

4 Methods

An initial build up of learning SQL, Pandas, and iPython notebook is needed. After learning how such a large data set gets utilized, there will be some data wrangling and cleaning of the set. Null values and outliers potentially have to be removed. Exploratory data analysis and visualizations will be created, including histograms, box plots, and possibly tree maps if Pandas supports that visualization.

After that, a some more detailed analysis will begin. Regression models and hypothesis testing will be conducted to determine the significance of the regression and connections between variables. Other statistical learning techniques, including classification trees and support vector classifiers, will be discussed to determine which methods can be best used to predict what products a customer will purchase next, along with the other questions to be answered.

5 Deliverables

Deliverables to be included are:

- An iPython notebook, posted to github, with code of visualizations and data analysis techniques.
- A three to five page technical paper documenting the questions, data set, methods, and conclusions of the project.
- A slide deck, between ten to twenty slides, of the project that could be used at a presentation during a job interview.