Capstone Project One Proposal

Title: Predicting the shopping orders from previous history.

Problem: Businesses would like to know what customers are going to buy in their future purchases. This knowledge will assist businesses in managing their supply chains to reduce excess inventory, saving cost, and ensuring products are on hand for customers, boasting customer satisfaction.

Who might care? Retail businesses, supply chain managers, product marketing agencies, product producers, and individual customers. The business itself will be better able to manage inventory levels, which could result in substantial cost savings by not over stocking merchandise. Products, perishable or fashion dependent, will be fresher or more in vogue and therefore more likely to sell. Customers could benefit from a recommender that before check out made recommendations that the customer had meant to buy but forgot. Market agencies could adopt the recommender to find potential undiscovered customers whose shopping habits favor their product. Ultimately, it is a tool that a business can use in supply management and sales boosting.

Data: The data will be obtained from InstaCart. This data contains purchase histories for 4,000 customers with orders from 4-200 per customer. Each order lists the sequence items were bought in, the aisle and department each item is found in, the time of day and day of the week for each order, and days from previous order.

Modeling approach: Data is spread across six csv files which will need to be combined (data wrangling). It is fairly clean but we will need to look for cleaning opportunities and imbalances. We want to predict the likelihood of an item being in the users next

basket. We plan to try several different recommendation algorithms and choose the best performer. Initially, a churn approach for items previously in a users basket will be used to predict item retention. We would like to couple this with a similar basket assessment across all user baskets to provide probabilities that items are bought together in an attempt to identify new items that might appear. A F1 score maximization seems appropriate for deciding on best recommender.

Possible limitations: The prediction of the model will likely never be 100%, therefore, attention will have to be paid to how to determine what constitutes a good recommendation. The imbalance of item reorders and the number of total user orders will add a challenge to analysis.

Deliverables:

1. Codes (notebooks) for:
    a. Data wrangling and cleaning
    b. Exploratory analysis
    c. Machine Learning model

2. Report on the capstone project

3. Presentation on the capstone project

4. GitHub repository