Project Proposal

1. The Problem and The Client

“Creating a Grocery Product Recommender for Instacart”

In the eCommerce shopping experience product recommendations come in many forms: they may be used to recommend other products on one product’s page (Amazon’s “Frequently bought together” feature for instance) or they may be used on the checkout page to show customers products they may be interested in based on their total order.

Instacart is an online grocery delivery service that allows users to place grocery orders through their website or app which are then fulfilled and delivered by a personal shopper- very similar to Uber Eats but for grocery stores. In 2017 they released a year of their data composed of about 3.3 million orders from about 200,000 customers.

Through the machine learning model and EDA, my best hope is to answer questions like:

- What product will the user buy again?
- What product will the user try for the first time or add to the cart during a session?
- What product is most likely to be ordered late at night?

Further, recommendations may be more helpful if they are targeted towards a specific segment of customers, rather than made uniformly. For instance, if one group of customers tends to buy a lot of non-dairy milk substitutes and another group tends to buy traditional milk, it may make sense to make different recommendations to go along with that box of Cheerios. In order to make tailored recommendations, Instacart users must be segmented based on their purchase history using K-Means clustering and then made recommenders based on the product association rules within those clusters.

2. The Data

The dataset is a relational set of files describing customers' orders over time. The goal is to predict which products will be in a user's next order. The dataset is anonymized and contains a sample of over 3 million grocery orders from more than 200,000 Instacart users. For each user, it is provided between 4 and 100 of their orders, with the sequence of products purchased in each order. We also provide the week and hour of the day the order was placed and a relative measure of time between orders.

Dataset: https://www.kaggle.com/c/instacart-market-basket-analysis/data

Each entity (customer, product, order, aisle, etc.) has an associated unique id. Most of the files and variable names should be self-explanatory.

- aisles.csv
- departments.csv
- order_products__*.csv

  These files specify which products were purchased in each order. order_products__prior.csv contains previous order contents for all customers. 'reordered' indicates that the customer has a previous order that contains the product. Note that some orders will have no reordered items. You may predict an explicit 'None' value for orders with no reordered items.

- orders.csv

  This file tells to which set (prior, train, test) an order belongs. You are predicting reordered items only for the test set orders. 'order_dow' is the day of the week.

- products.csv
- sample_submission.csv

3. The Deliverables

The final deliverables would be

- Jupyter notebook code file with a recommender system built for the above-mentioned model,
- Descriptive project report and a project presentation.