**Project Proposal for Capstone 2**

**Background/Context:**

Online grocers like Instacart provide a certain set of benefits to their end customers. For customers that are willing to pay a small fee, online grocers are able to shop for them based on their grocery requirements and deliver the groceries to them within a reasonable timeframe. End customers save the time spent doing groceries themselves.

Online grocers offer a software platform (both on mobile and desktops) that customers can use to browse various products/grocery items offered by different grocery stores that the online grocer has partnered with. Once the customer submits the order via the software platform, online grocers take care of that order via shoppers and drivers that they have employed or contracted with. For the service they provide, they take a fee from the customers and their partner grocery stores.

End clients that I have in mind for capstone 2 project are online grocers who value recommending relevant products to their customers. End customers would benefit by being aware of products potentially relevant to them and as a result, online grocers would benefit by potential upside to their revenue if the recommended products are purchased by customers.

**Problem to be solved and beneficiary of solution:**

Specific problem I will be addressing for online grocers is that of making product recommendations to customers based on their similarity to other customers. The primary beneficiary of solution to this problem would be marketing/advertising managers at online grocers who can potentially use this to increase revenue from targeted customers. From customer’s perspective, they would be made aware of products being purchased by customers similar to them in purchasing behaviors

**Data gathering:**

I will be using data provided at Kaggle:

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

The data is comprised of 6 files:

i) df\_orders (about 3M rows) : contains information regarding orders (users, orders, day of week order was made, hour of day order was made, days since prior order)

ii) df\_order\_products\_prior (about 30M rows): contains information regarding prior orders (orders, products, add to cart order, whether a product was reordered or not)

iii) df\_order\_products\_train (about 1M rows): contains information regarding orders to be used for training (orders, products, add to cart order, whether a product was reordered or not)

iv) df\_products (about 50K rows): contains information regarding products offered (products, aisle, department)

v) df\_aisles (about 100rows) : contains information regarding aisles that products belong to (aisles)

vi) df\_departments (about 20 rows): contains information regarding departments that products belong to (departments)

**Method(s) to be used:**

Data obtained from Kaggle would be explored to understand the users, their behaviors, their orders and their product preferences better. It would be cleaned and processed as necessary to prepare it for subsequent steps.

After data cleaning, features that represent the shopping behaviors of users would be selected and used for clustering/segmenting the users. The primary clustering method I plan to use is KMeans clustering method. Other clustering methods could be explored as well depending on time constraints:

1. DBSCAN
2. Spectral Clustering

Any user will have multiple orders. There are user behaviors associated with any specific order as identified by day of week order was submitted, hour of day order was submitted and days since prior order. User clustering will be performed in one of two following ways:

1. Start by clustering orders first and then assigning the most frequent/dominant cluster for that user to that user.
2. Start by taking the average of user behaviors first. Once that is obtained for each user, cluster users based on those average values

After clustering users using above two approaches:

1. Identify products that are popular within a given cluster
2. For each user in a given cluster, identify and recommend products from the above set of popular products that the user hasn’t purchased before.

**Deliverables:**

The deliverables would be a slide deck (powerpoint format), a written report ( word format) and python code (jupyter notebook format).