Capstone Project Proposal

Instacart Market Basket Analysis

Anushree Srinivas

Mentor: Srdjan Santic

Problem to be solved and Motivation:

Instacart, a grocery ordering and delivery app, aims to make it easy to fill your refrigerator and pantry with your personal favorites and staples when you need them. After selecting products through the Instacart app, personal shoppers review your order and do the in-store shopping and delivery for you. Currently they use transactional data to develop models that predict which products a user will buy again, try for the first time, or add to their cart next during a session. Instacart is now looking find a model that predicts the user's next order based on the previously purchased products. The ability to identify which products the customers are likely to purchase again, and automatically adding those to cart through obtained predictions or provide a seamless interface for doing so will enhance their user experience. These predictions of the products the customers are likely to order can be send in personalized communications to customers reminding them to order again, by highlighting the predicted products in those communications.

Client:

Instacart is looking to use this analysis to better serve their customers. The data science team at Instacart will be the client for which the conducted data analysis as part of the capstone project will be beneficial.

Data:

The dataset for this project is a relational set of files describing customers' orders over time. The dataset is anonymized and contains a sample of over 3 million grocery orders from more than 200,000 Instacart users. For each user, information related to between 4 and 100 of their orders, with the sequence of products purchased in each order are provided. The week and hour of day the order was placed, and a relative measure of time between orders is also available. The data will be obtained from the ongoing Kaggle Competition.

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

Approach:

Initially, I am going to dive deep into the data to get a thorough understanding of the different variables. I plan to do some data wrangling and cleaning to deal with missing values etc. Once the data is cleaned, I am going to perform some exploratory analysis to recover some interesting insights from the data. Feature engineering will also be applied to discover some new meaningful features that can be used while building the models. Finally, different models will be built using machine learning algorithms. Data visualizations will help in communicating these insights and tell a story.

Deliverables:

The capstone project will produce a Presentation (slide deck), a report (highlighting the approach, findings etc.) and the Python Code in the form of a iPython Notebook that will be submitted into my Github repository. A final paper explaining the problem approach, findings and recommendations will also be submitted.