IST 718: Project proposal

Title:

Instacart Market Basket Analysis

Group Members:

Akshit Saliyan

Bhaumik Doshi

Palak Nyati

Radhika Bhange

Introduction:

This project focuses on the Market Analysis, an area which deals with customers on a regular basis and strives to understand customer behavior based on their purchases. When a customer shops from meticulously planned grocery lists, their unique food rituals define who they are. Instacart is a grocery ordering and delivery app. After selecting products through the Instacart app, personal shoppers review the customer's order and do the in-store shopping and delivery for them.

Currently Instacart is using 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. Using this open-sourced data of 3 million customer orders over time we are predicting which previously purchased products will be in a user's next order.

Dataset:

The dataset for this competition 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. The week and hour of day the order was placed, and a relative measure of time between orders are also a part of the dataset.

Orders: 3.4m rows, 206,000 users
Products: 50,000 different kinds of products

Goal: Identifying customer buying habits and building a recommendation system.

Task: Determining association among different items that customers buy.

Expected Result: By identifying association between different items that customers tend to buy, it can help develop marketing strategy by gaining insight into which items are frequently purchased together by customers.

Expected Problem: Simply applying the algorithm on the available data will generate profusion of rules so we need to pick the rules carefully so that we don't lose any valuable information and have a strong recommendation system.

Goal: Identifying best customers.

Task: Determining how recently and how frequently a customer has bought.

Expected Result: By identifying the recency and frequency, we can determine quantitatively which customers are the best.

Expected Problem: The dataset does not have the amount each customer has spent in each of his/ her order which can affect the prediction of the best customers.

Goal: Identifying the items that a new customer might buy.

Task: Determining the behaviour of order pattern for new customer in the dataset

Expected Result: Understanding and predicting the items can be bought by the new customers

Expected Problem: The dataset might not have sufficient data for the new customer's first order

Goal: Understanding which products are reordered and when.

Task: Calculating reorder ratio for different departments and time of the day.

Expected Result: Understanding when and which section of items are more likely to be reordered helps in optimizing inventory management.

Expected Problem:

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

We plan to use Databricks, Jupyter Notebooks, Spark, Hadoop.

1. Classification: KSVM
2. Customer Segmentation using PCA (Principal Component Analysis), Clustering
3. Association Rule Mining

We plan to use Matplotlib, geoplotlib, seaborn to generate data visualizations throughout our analysis.