Instacart Market Basket Exploratory Analysis

Springboard Capstone Project-I Proposal By Megha Saini

Problem

Instacart recently announced its first public dataset, "The Instacart Online Grocery Shopping Dataset 2017". It is an anonymized dataset and I plan to use it for the Exploratory Data Analysis (EDA) by using a combination of inferential statistics and data visualization techniques to find interesting trends and identify significant features in the data set.

Target Audience

Data science techniques required to accomplish this task is of primary interest to Instacart, an American company that operates as a same-day grocery delivery service. At Instacart, customers select groceries from various retailers through a web application and a personal shopper delivers the goods same-day. These personal shoppers are independent contractors or part-time employees looking to make some extra income through Instacart besides their primary job.

Insights derived in this project will enable Instacart gather invaluable insights about customer purchase behavior, product portfolio and basket analysis. It will also help part-time Instacart shoppers to understand the optimal days and optimal hours to work based on the prime time for maximum orders placed by Instacart users/customers.

Furthermore, there is an increasing number of online-grocery companies like Fresh Direct, Peapod, etc. and traditional grocery stores that offer online grocery shopping for its shoppers. It is reasonable to think that many of them have recorded data of purchase orders over many years. Development of data science techniques in this project will potentially benefit these online grocery shopping companies to derive insights for analysis on their own datasets.

Data Acquisition

This anonymized data is released by Instacart and is available for download from their **website**. It contains a sample of over 3 million grocery orders from more than 200,000 Instacart users. For each user, they provided between 4 to 100 of their orders, with the sequence of products purchased in each order. The dataset also offers information about the week and hour of the day the order was placed, and a relative measure of time between orders.

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

a) *orders.csv* - This file provides a list of all the orders (order_id) in the dataset. Each row represents an order. For each order, it provides information on user (user_id), hour and day of the week when order is placed (order_hour_of_day, order_dow), and days since prior order (days_since_prior_order)

b) *products.csv*- This file contains names of the products with their corresponding product_id. It also includes the aisle id and department id.

c) *aisles.csv*- This file lists the aisle id and name

d) *departments.csv*- This file lists the department id and name

e) *order_products_*.csv*- These files represent the training and prior datasets. These provide information on order id (order_id), products ordered (product_id), the order in which the products were put into the cart (add_to_cart_order) and if a product is a re-order (1) or not (0) (reordered). Some orders will have no reordered items.

Methodology

Insights derived in this project will be based on inferential statistics, data wrangling and data visualization techniques to derive interesting patterns, trends and identify significant products for reorders. Further analysis can also be done on this dataset to predict products that a user will buy again, try for the first time or add to cart during a session. Exact modeling approach is yet to be defined for latter part of the analysis.

Deliverables

The source code will be shared on a public GitHub repository. In that repository, a final paper will be made available explaining the problem, EDA techniques and a short discussion of the results. A slide deck used in the presentation of this project will also be available for download along with the source code.