Instakart Market Basket Analysis

Name : Shreyansh Singh

Date: 09/Feb/2018

NUID: 001835018

Introduction:

Instakart is a grocery shopping store that caters to consumer needs through online deliveries. Post deliveries the customer review the services provided. Based on the orders and reviews, Instakart have provided their data for various analysis purposes. Many a times the online stores seek the help of data scientists to help them to increase their sales and performance.

Background Research:

Here we will do analysis around the products and customers like what are the favorites of a particular customer, or which products have contributed the most towards overall revenue, or which are the days that have attracted most number of customers or which time is most favorable for any part time employee to work so that it gets maximum pay, etc.

We will be using 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, we provide between 4 and 100 of their orders, with the sequence of products purchased in each order. We also provide the week and hour of day the order was placed, and a relative measure of time between orders.

Data Sources:

We have used dataset available on Kaggle:
https://www.kaggle.com/c/instacart-market-basket-analysis/data

1. Aisles.csv
2. Departments.csv
3. Order_Products_Prior.csv
4. Order_Products_Train.csv
5. Orders.csv
6. Products.csv
7. Sample_Submission.csv

Algorithms Used:

1. Random Forest:

   A random forest is a meta estimator that fits several decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement if bootstrap=True (default).

2. Exploratory data analysis:

   Exploratory data analysis helps one understand the data, to form and change new theories, and decide which techniques are appropriate for analysis. After a model is finished, exploratory data analysis can look for patterns in these data that may have been missed by the original hypothesis tests. Successful exploratory analyses help the researcher modify theories and refine the analysis

3. Logistic regression:

   Logistic regression, or logit regression, or is a regression model where the outcome variable is categorical. Often this is used when the variable is binary (e.g. yes/no, survived/dead, pass/fail, etc.). Logistic regression measures the relationship between the categorical response variable and one or more predictor variables by estimating probabilities.

4. Naive Bayes:

   Naive Bayes classifier are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

5. Clustering:

   Clustering the process of organizing objects into groups whose members are similar in some way. A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

References:

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

https://www.kaggle.com/vamsikrishna/exploratory-analysis-instacart

https://www.kaggle.com/sudalairajkumar/simple-exploration-notebook-instacart

http://blog.kaggle.com/2015/06/19/facebook-iv-winners-interview-2nd-place-kiri-nicholaka-small-yellow-duck/

https://www.kaggle.com/asindico/customer-segments-with-pca