CS F415- Project Proposal for Implementation/Development based Project

Team Members:

Sai Ashwin Kumar C - 2019A4PS0628H
Hari Krishna Dhamodaran - 2019A4PS1263H
Venkata Sai Adarsh Kone - 2019A3PS0388H
Jayant Lingamaneni - 2019A7PS0005H

Title: Data Mining for Product Marketing Analytics

Abstract:

The aim of this project is to perform data mining and analysis on a huge dataset of product orders of a brand's website and suggest marketing strategies based on the analysis. We would also like to obtain a lot of interesting inferences and patterns from the same, for understanding the behaviour of the users for planned and cost-effective marketing. We plan to achieve this by employing data visualization features in pandas, associative rule mining, frequent pattern mining using apriori rule, Clustering using K-means, Classification using logistic regression and random forest method, and many more methods if required. We would also try to extend the same problem to another case like social media marketing.

Problem Definition:

The dataset is from an online store of sporting goods: clothing, shoes, accessories and sports nutrition. On the main page of the store, they show users banners in order to stimulate their sales. Now one of 5 banners is randomly displayed there. Each banner advertises a specific product or the entire company. The experience with banners can vary by segment, and their effectiveness may depend on the characteristics of user behaviour. Now the job is to test the effectiveness of these banners from the available data and strategically decide when and what should the banners show.

State-of-the-art:

There are various Machine Learning methods employed in the field of predictive business analytics. Classification methods and algorithms like logistic regression, KNN, SVM, decision trees, Bayesian networks are widely used for the segmentation of customers into different classes. Clustering is also a widely used process of dividing an organisation's customers into groups or 'clusters' that reflect similarity amongst customers in that particular group, the difference from the former being that it is an unsupervised learning method that doesn't come with any class labels. Packages like pandas, matplotlib, seaborn are also powerful tools for data visualisation in Python.

Novelty:

The previous works performed on the dataset, as obtained from Kaggle, don't contain many aspects of association analysis, rather they are more focussed on the segmentation part, which would like to work on. Further, we also aim to apply ensemble/classifier combination methods that improve classification accuracy by aggregating the predictions of multiple classifiers.

Dataset:
We have information about which banner appeared to the user, whether he/she clicked on it, as well as information about user purchases. The size of the dataset is around 1.2 GB with as much as 8.5 million instances of data. The link for the source from it was taken:

https://www.kaggle.com/podsyp/how-to-do-product-analytics

- order_id - unique purchase number (NA for banner clicks and impressions)
- user_id - unique identifier of the client
- page_id - unique page number for event bundle (NA for purchases)
- product - banner/purchase product
- site_version - version of the site (mobile or desktop)
- time - time of the action
- title - type of event (show, click or purchase)
- target - target class
- State - State of US from which order was made

We would like to first simplify the data so that it could be processed in a better way for further analysis. Then we can apply different classification (using ensemble method) and clustering techniques for customer segmentation based on various parameters and establish patterns and correlations (for example, relating the time of user activity, the action of the user, season of the year (from date), product together) using mining methods. Particularly we perform the 'Market Basket analysis', used by retailers to increase sales by better understanding customer purchasing patterns.