*Thaiyalnayagi Karthik*
*Data 606 - Capstone Project – Part II*
*Spring 2020*

**Predicting Instacart Customers Purchasing Behaviors**
**(When they will make their next purchase and what products to expect in that purchase)**

**Literature Review:**

Increasing competition in the market along with rapidly changing customer purchasing behaviors urges companies to invest money and time on analyzing customer's purchasing behavior to enhance their shopping experience with the company, thus satisfying and retaining customers, and sustain and grow over others in the market [1]. Most of the companies in this research focus on predicting customer's purchase intentions (customers' next purchase and possible products in their next purchase or any new product to be added in their purchase), using customer's purchase history [2]. Answers to these questions can help them be prepared (i.e., stock replenishment) to satisfy customers' needs instantly without any delay [3] and build recommendation systems to make their shopping easier and interesting. For instance, Instacart is using the XG-Boost algorithm to predict item-based availability to increase customer experience without making them disappointing when they look for a product [4]. Moreover, the complexity of these researches or predictions lies in how efficiently and accurately their model can predict customer purchase behavior. Decisions took based on incompetent analysis/models can lead to a huge loss. So, companies are still working on improving their models to predict precisely. On the other hand, individual researchers are trying to build efficient models for predicting consumer's purchasing behaviors based on real-time datasets using machine learning algorithms. The most commonly used algorithms for next-purchase prediction are XG-Boost, Naïve-Bayes, gradient tree boosting, Random Forest [5][3][6]. Whereas for multi-label classification problems like product preference prediction, transformed Logistic Regression [7], transformed Naïve-Bayes, adapted Multi-label KNN [8], convolutional neural networks [9], mostly used. As these models applied to different datasets, the results of these are not comparable directly. Further, applying different tuning techniques can lead to different results. Hence in this project, different models will be created and tested to find the best model for both consumer's next-purchase and consumer's product preferences in their next purchase separately.

**Data Cleaning:**

Datasets are almost clean, except for some missing values on the department, and aisle for some product names which were filled with similar values matching with products and remaining were categorized into others.

**Exploratory Data Analysis Findings:**

The combination of six datasets gives information on product details, product added sequence, whether it is reordered or not, hour and day of the week, days since prior order, aisle, and department of the products of about 3.4 million orders made by 0.2 million users. For every user, 3 to 99 histories of purchases are given. Most of the orders are loaded with less than 20 products. However, outliers containing more than 100 products in an order is also seen. Further, among 49513 ordered products, 60 percent were reordered. Overall, the key take-over from the explorations are, $0^{th}$ and $1^{st}$ day of week (days anonymized) high orders are placed whereas 9 am to 4 pm all days is the peak sales time; A high number of orders seems to be placed on a weekly, biweekly or monthly basis; Fresh fruits seem to be the topmost preference of users followed by packaged and fresh vegetables. Further, we could see a direct relationship between the sequence of products added to the most reordered products, i.e., initially added products mostly reordered. Finally, the distribution of product occurrences in the dataset is uneven, with many being repeated only 4 to 10 times while few ordered more than 20,000 times. Hence it needs to be considered while designing the model.

**References:**

1. Nikhil Agarwal, “How Amazon, Flipkart use data analytics to predict what you are going to buy”, 2018. Available: https://www.livemint.com/Companies/RX5eOy12n5JFJu617G5GnM/Amazon-Flipkart-data-analytics-ecommerce.html
2. Instacart, “Instacart Market Basket Analysis”, 2017. Available: https://www.kaggle.com/c/instacart-market-basket-analysis
3. Andrés Martínez, Claudia Schmuck, Sergiy Pereverzyev, Clemens Pirker, Markus Haltmeier, “A machine learning framework for customer purchase prediction in the non-contractual setting”, 2020. Available: https://www.sciencedirect.com/science/article/abs/pii/S0377221718303370
4. Abhay power, “Predicting the real-time availability of 200 million grocery items”, 2018. Available: https://tech.instacart.com/predicting-real-time-availability-of-200-million-grocery-items-in-us-canada-stores-61f43a16eafe
5. Baris Karaman, “Predicting Next Purchase Day”, 2019. Available: https://towardsdatascience.com/predicting-next-purchase-day-15fae5548027
6. Featuretools, “Predict Next Purchase”, 2019. Available: https://www.featuretools.com/project/predict-next-purchase/
7. Prateek Joshi, “Predicting Movie Genres using NLP – An Awesome Introduction to Multi-Label Classification”, 2019. Available: https://www.analyticsvidhya.com/blog/2019/04/predicting-movie-genres-nlp-multi-label-classification/
8. Kartik Nooney, “Deep dive into multi-label classification...! (With detailed Case Study)”, 2018. Available: https://towardsdatascience.com/journey-to-the-center-of-multi-label-classification-384c40229bff
9. Guanglei Zhang, Lei Chen , Yongsheng Ding , “A multi-label classification model using convolutional netural networks”, 2017. Available: https://ieeexplore.ieee.org/document/7978871