# Predicting Instacart Customers Purchasing Behaviors

**(When they will make their next purchase and what products to expect in that purchase)**

**Problem Definition:**

Customers are the heart of any business. Especially for companies like Instacart (mobile and web-based on-demand grocery delivery company), dependency on customers is too high. Their success entirely relies on customer experience and customer satisfaction. So, focus on customer behavior will help them keep track of their business and make strategic decisions to improve business performance from the customer perspective. Further, customer purchase behavior analysis will also help understanding customers’ intention of further buying and loyalty to the company, which is crucial for business success [1]. This project aligns with this problem statement and will try to bring out valuable insights about customer’s purchase behavior.

# Project Objective:

This project aims to analyze the Instacart purchase orders, understand customers’ purchasing behavior based on their purchase history and predict when a customer will make their next order and what are the products, they might have in their next purchase order.

# Dataset Description:

For this project, I am going to make use of the Instacart purchase order dataset from the INSTACART website [2]. This dataset is a collection of six datasets comprising around 3 million order details, including the information about the user who made the order (UserID), the list of products in each order (ProductID), department, and aisles to which the product belongs. On the other hand, to protect the privacy of their users and their retail partners, Instacart had shared only an anonymized version of the purchase orders [3]. So, this dataset can be used freely without requiring any further processing to protect user privacy.

**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 [[4]](https://www.livemint.com/Companies/RX5eOy12n5JFJu617G5GnM/Amazon-Flipkart-data-analytics-ecommerce.html).  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 [5].  Answers to these questions can help them be prepared (i.e., stock replenishment) to satisfy customers’ needs instantly without any delay [6] 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 [7]. 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 [8][6][9].  Whereas for multi-label classification problems like product preference prediction, transformed Logistic Regression [10], transformed Naïve-Bayes, adapted Multi-label KNN [11], convolutional neural networks [12], 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 for which I filled the missing data with similar values, and I categorized remaining 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 (Figure 1). 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 (Figure 2). Overall, the key take-over from the explorations are, 0th and 1st 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 (Figure 3); 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 (Figure 4). 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.

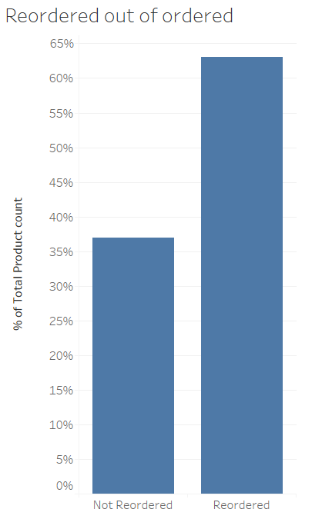
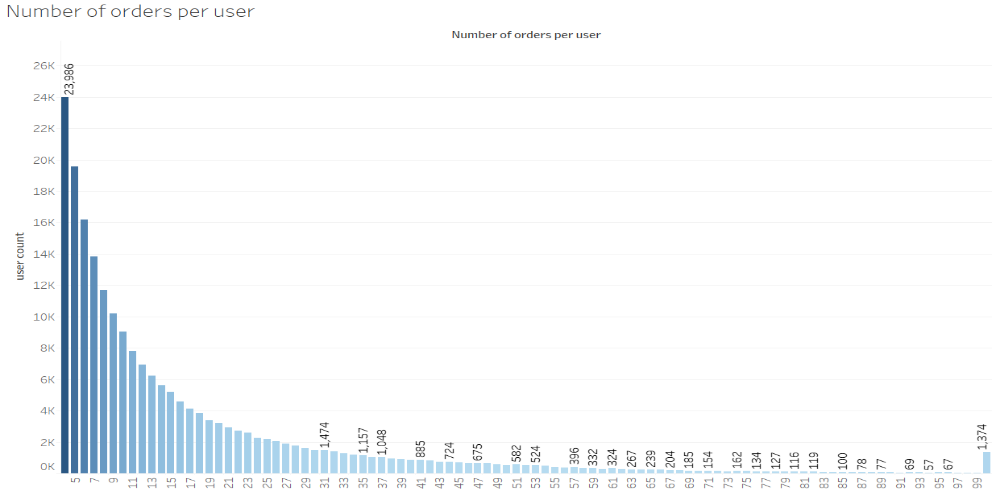


Figure 1: Number of orders per user Figure 2: Reordered product ratio

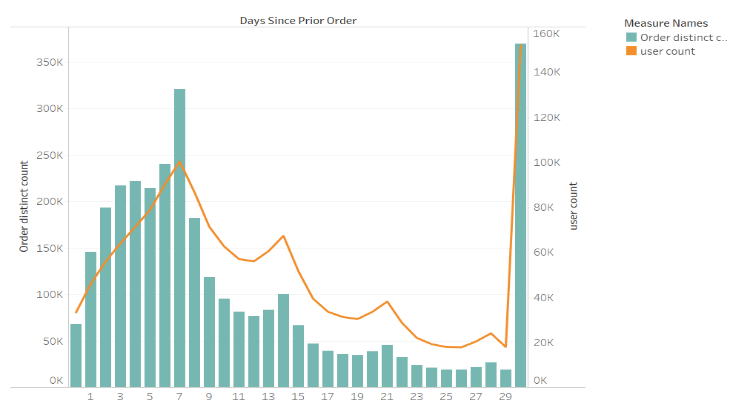
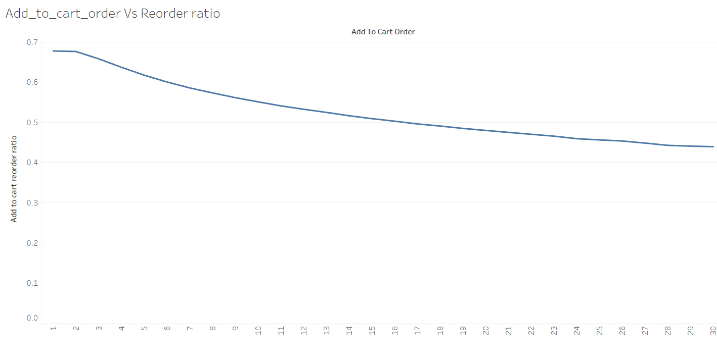
 

Figure 3: Customer Order frequency Figure 4: Add to cart versus reorder ratio

# Approaches to achieving project goals:

# This project deals with two predictions namely when the user will make the next purchase (prediction-1) and what are all the possible products in the next purchase order (prediction-2). In the case of prediction 1, concerning exploratory analysis findings (Figure 3), I categorized the customers’ next purchase interval into two classes (Class 0 – less than 15 days since prior order and Class1 – greater than 15 days but less than 30 days since prior order). For prediction 2, to simplify the multi-label classification problem, I reframed the question as will this product be in the user’s next order or not, turning the problem into a binary classification problem. Further, I will be predicting only the products which a customer had reordered at least once.

# Assumptions on the model:

# Before building the model, I made some assumptions to stick to the project goal. The assumptions include, all the customers in the model will surely make another purchase, that too within 30 days. This assumption is based on the inference from the exploratory analysis, which showcases that the customers in the dataset tend to make an order within 30 days. Hence, customer churn is not handled in this model. Further, as this model is built with the customers who made at least four orders, the prediction on new customers is not assured.

# Data Preprocessing & Feature Engineering:

# With been defined the approach for modeling, I extracted various features including user-product features, such as purchase frequency, reorder rate, average add to cart order, and so on from the history of the users’ orders based on data exploration findings. Further, to increase the complexity, have also added user’s features such as overall purchase frequency, average cart size, etc. Similarly, I extracted overall product features to add as the input to the model. After extracting the features, I converted the categorical columns to the corresponding numerical category using Label Encoder and performed log transformation on a few columns such as total orders and visit duration for which the range is broad. Then, split the data into train and test set for building the model.

# Modeling overview:

# For prediction 1, several classification algorithms were trained and tested. Among those, the XGBoosting Classifier performed better compared to others producing a test accuracy and f-measure of 70% each on both the classes. I tried adding some features and removing some outliers like customers making orders on the same day or in one or two days due to a few missed products in their immediate past purchase. However, they did not seem to boost accuracy. Hence, I move forward with the previously extracted features itself. Then I extracted the important features using feature importance as such that the model becomes simple but not compromising on the accuracy. Further, I tuned the model with parameters such as scale\_pos\_weight and max\_depth to handle the prediction on minority class labels in the model. Finally, to check the robustness of the model, I evaluated the model with the cross validator and ended up with 69.85% mean accuracy with a standard deviation of 0.49%. Later, I analyzed the prediction result over the true labels to find some way to improve the accuracy, I found that though all the customers tend to buy within 30 days, many of the customers did not follow a definite pattern of buying weekly or monthly basis.

# On the other hand, even for prediction 2 XGBoosting Classifier seems to perform better compared to other algorithms. The model was able to correctly identify 76% of the not buying products over around 60% of the products going to be bought. The scale\_pos\_weight parameter of the XGBoosting Classifier is adjusted for the model to generalize the features rather than working biased on majority class labels. Moreover, the model provided an overall accuracy of 71.34% with 0.15% standard deviation and 80% and 51% F-measure on Non-Buying and buying product classes respectively.

# Model Result Interpretation:

# 

# Figure 5: (a) Buying a product prediction (left) and (b) Not buying a product prediction (right)

# Though the model does not seem to correctly identify products to be purchased at greater accuracy, I found that plotting the true labels based on the department makes sense. That is, as you can see in Figure 5(a), for few departments such as department id 3, 4, 7, 16, 20, the model had predicted at a rate of more or around 70% accuracy. So, I recommend implementing this model in the above-mentioned departments would help increase the efficiency of the stock management of the company. On the other hand, Figure 5(b) clearly shows that the model is also performing well in identifying the products that are not going to be purchased. So, I hope that implementing this model will surely show some improvement in the stock management of the company reducing spending money and space on unnecessary products.

# Conclusion and future work:

# Overall, this project studied the behavior pattern of the Instacart customers and predicted the customers’ next purchase interval and product choices in their next purchase based on their previous purchases with around 70% accuracy. Further, the model is designed in such a way that prediction on buying products over non-buying products is balanced for better stock management decisions. Moreover, as mentioned above, implementing this model in the departments on which the model is performing better would give good results.

# Some of the next steps to improve the model would be performing transfer learning or clustering the customers based on their product preferences and building a separate model for each cluster. Though the improvement is not quite assured, there is a possibility to get a different or better result.

**References:**

1. Ben, “The Importance of Customer Satisfaction”, 2017. Available: https://www.thoughtshift.co.uk/the-importance-of-customer-satisfaction/
2. Instacart, “The Instacart Online Grocery Shopping Dataset”, 2017. Available: https://www.instacart.com/datasets/grocery-shopping-2017 [Accessed: Jan 2, 2020]
3. Jeremy Stanley, “3 Million Instacart Orders, Open Sourced”, 2017. Available: <https://tech.instacart.com/3-million-instacart-orders-open-sourced-d40d29ead6f2>
4. 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>
5. Instacart, “Instacart Market Basket Analysis”, 2017.  Available: <https://www.kaggle.com/c/instacart-market-basket-analysis>
6. 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>
7. 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>
8. Baris Karaman, “Predicting Next Purchase Day”, 2019.  Available: <https://towardsdatascience.com/predicting-next-purchase-day-15fae5548027>
9. Featuretools, “Predict Next Purchase”, 2019.  Available: <https://www.featuretools.com/project/predict-next-purchase/>
10. 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/>
11. 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>
12. Guanglei Zhang , Lei Chen , Yongsheng Ding , “A multi-label classification model using convolutional netural networks”, 2017.  Available: <https://ieeexplore.ieee.org/document/7978871>