Capstone 2 Project Milestone

by Gerald Wijaya

Introduction

Unlike other ecommerce sites, Instacart holds no inventory of products. Yet users are able to shop their way through the application for groceries and other essentials that can be found in supermarket. An interesting fact is that the company creates their shopping site as a way to place order with products, claiming to come faster than the shipping from a warehouse like Amazon. How is this possible? Well the idea is not that hard to think of. At the backend after each order is placed, a program sends out errand notification to their couriers to do the shopping in the grocery store of their vicinity for the customers and deliver the order to their house. The string attached to this application is an annual subscription fee in the expense of the customer's convenience. The reason the company does not need to hold any inventory is the intervention of a middleman to do your bidding by retrieving the stock of products from a supermarket near the customers.

Problem Statement

Instacart aims to create a dependency towards their system for customers who likes their groceries in their home delivered. The challenge with the service is tracking the accurate amount of inventory from a local store near the courier and delivery within the time window. Doing so will boost the loyalty of customer and hence increase the total revenue the company can make from the annual subscription. The goal at the end of this project is to be able to predict the customers' next order, knowing their historical order. Being able to create a feature element on the app that shows accurate prediction of customers order, could enhance the Instacart users experience and help them order faster than their previous orders. This in turn also increase the findings of inventory needed around the area, perhaps even partnering with local supermarket on controlling the supply chain of certain products in the area. Therefore, being able to get the accuracy about 0.3 from the prediction model is the key of this project.

Data Knowledge

Instacart has submitted 3 million rows of dataset for the public to take on the challenge of predicting the reorder ratio of a specific product on their 200 thousand anonymized customers' next visit. There are seven datasets provided from the Kaggle challenge set; aisles, departments, orders, orders_product_prior, orders_product_train, products, and sample_submission.

“The dataset for this competition is a relational set of files describing customer's' orders over time. The goal of the competition is to predict which products will be in a user's next order. 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. For more information, see the blog post accompanying its public release.”

The following step after entering EDA is feature engineering. The datasets provided contain 15 unique variables, 70 features that can be used for modeling and applying Machine Learning to find the F1 score (measure of test accuracy). Ideally, the preference to solve this problem is to stick with less computationally intensive/loss costly and yet more accurate option. Let’s start with going through some of essential features from the simplest to the most intricate ones:

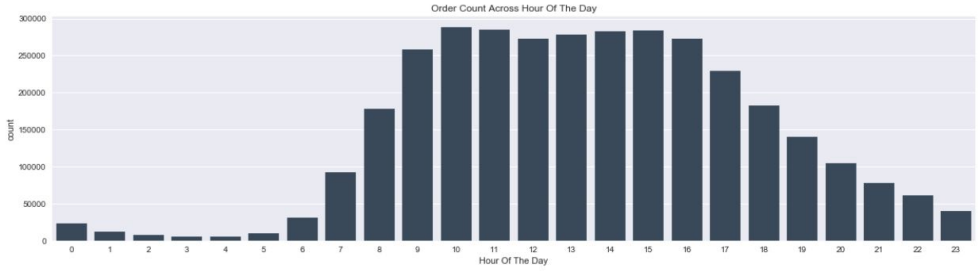
- Aisle: contains the ID of the aisle by numerical order and its name for each respective aisle
- Departments: contains the ID of the department by numerical order and its name for each respective departments.
- Products: contains ID of the products, including information of which aisle and department the product is located
- Orders: contains the ID of an order, following with the user ID that placed that particular order and a column called order number, which is the Nth number of order the user has placed. There is also columns showing time element such as orders from Day of Week, Hour of Day and number of days before the last order. Another prominent element of this feature is eval_set, which for this dataset contains both entries for training data (about 131k users) and testing data of about 75k users that needs to be predicted.
- Orders_Prior (Train and Test): There are two datasets given for this feature, one is for training and the other testing. The two datasets contains the same elements; the ID of the orders and users, following the add_to_cart_order, which put the position of the product in a list. Then one of the important element in this dataset is that of reordered value, which gives entries of only 0 or 1 depending on whether the product was a repeat order from before.

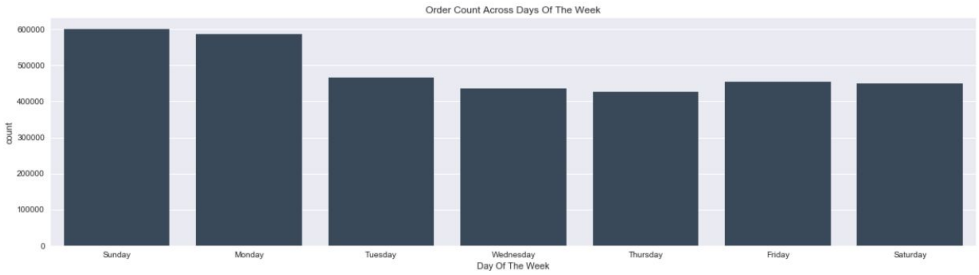
Initial Exploratory Data Analysis (EDA)

The analyses of the dataset consist of data from the orders data, products data, aisles data, department data and reorder data count data.

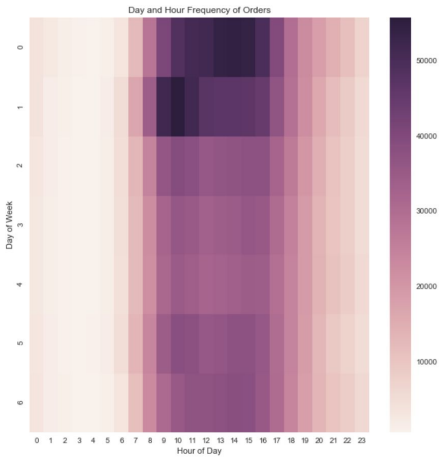
Orders Data:

From the plot below, orders data is shown relative to time in day, hours, and weeks. It can be depicted that the peak time for orders made by customers are between 10:00 to 15:00, possibly during the work times. It looks like orders are mostly made during the weekends and from the above, during the day time. Let's look at the results of hours and day further to find which day and time is the most frequent order coming from.

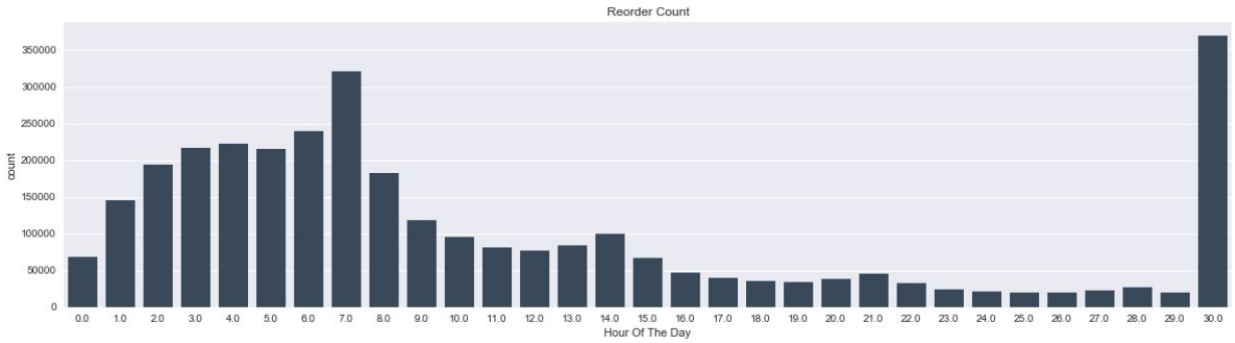




From the heatmap above, it appears that the darker contrast of color shows the most frequent orders made during the day and time of the week. Days of the week value of 0 and 1, which corresponds to Sunday at 2 pm, generally most orders are done between 10 am to 5 pm
and Monday at 10 pm, generally crowded between 8 am to 5pm



The trend understood from above is fairly interesting, as it seems like the end of weekend is when customers starts to think on what to purchase or chose to purchase and on Monday (regular workday), customers may have choose to come back to the order cart to finish the orders or place another order. This information can be analyzed further when looking at the reorder counts.

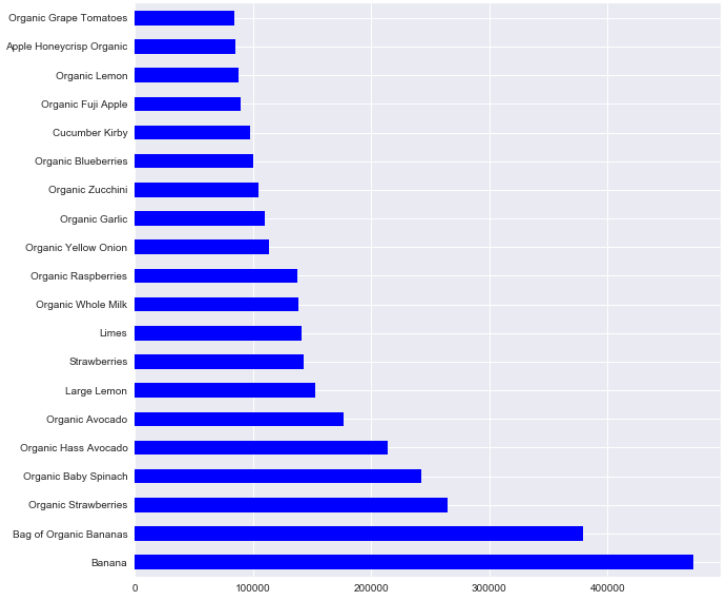


We want to know when customers will start working on the orders prior to the orders made, the reorder dataset can provide this visualization. From the chart above, it can be seen the reorder timeframe is about 30 days. The peaks of the days to reorder indeed have a pattern of every 7 days (7, 14, 21,28 with 30 days being the highest peak). This could mean that the customer is taking advantage of their 30 days trial for Instacart. Hence, a trend can be seen for customers shopping pattern; they tend to order about every 7 days.

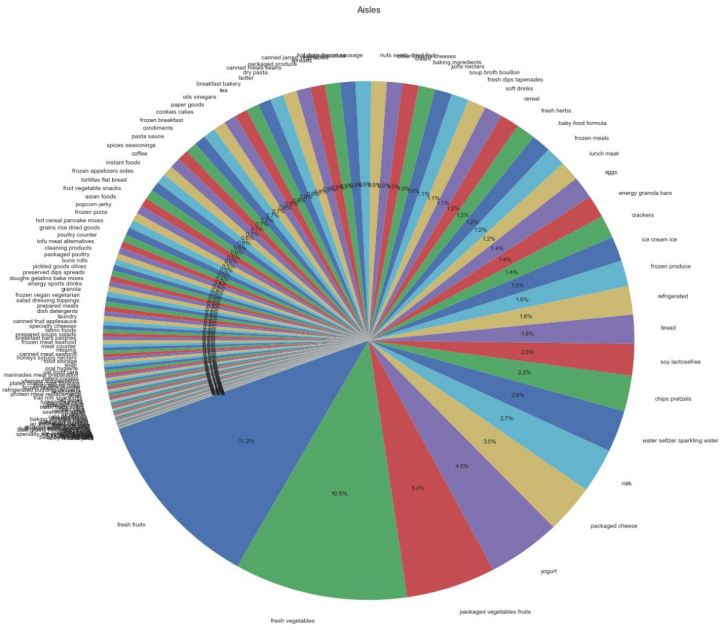
Generally, no cart orders made are less than 3 products and most cart are capped at 100 products per orders. This could mean that Instacart made a minimum amount or minimum price in each order made by customers.

Products Data:

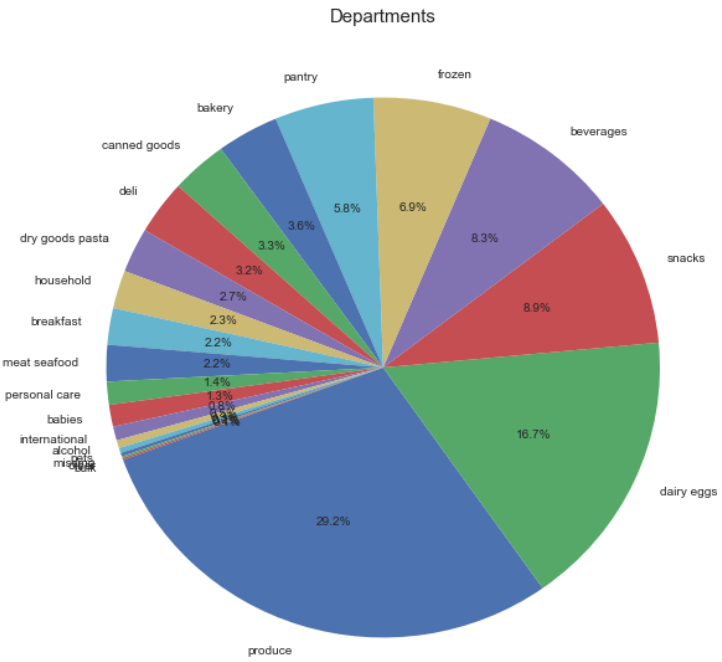
The products dataset analysis here focus on the compiled list of orders by customers, specifically on which items that users like to buy from Instacart. This is categorized in many different ways and can be analyzed in many different permutations or combinations of the categories of the products.



It is apparent that the top products purchase on prior orders are mostly healthy items, this could be the mission for Instacart to bring to their customers; Healthy products for at a regular price compared to in-store purchase. To further see whether such information is true, let's merge the dataset with the aisles to see more specifications.



From the depiction from the chart above, The most ordered products comes from the aisle: Fresh Fruit (11.2%), Fresh vegetables (10.5%), Packaged Vegetables Fruits (5.4%). This result seems usual since the most ordered products are vegetables/fruits or in general healthy products. Now let's look at the problem further by inserting the departments into consideration.



Similarly for the departments review, healthy products prevails in the option for customers who chose to shop online for grocery. This includes departments from produce (29.2%) and dairy eggs (16.7%). While fresh options tend to be more popular among the purchases, prior purchase on other proportion of orders may have come from habitual purchase. Habitual purchase are more likely group with other products during reorder based on how closely the associated products are. The next step after EDA would be feature engineering, which involves determining extra attributes by further developing the features of the dataset using these three categories: Product features, user features, and product-user features. The result will be shown in the final report.