Instacart Analysis Report

By: Parin Patel

Data Background:

Instacart is a same-day grocery delivery and pick-up service company that lets customers shop at their local grocery stores online or from their mobile phones. The service is known for its use of personal shoppers who then go retrieve the items from the grocery store and deliver it to the customers home. Valued at over $7.8 billion, Instacart is accessible in all 50 states, and over 5,500 cities in the US and Canada. In addition, the company has partnerships with over 300 retail patterns, resulting in access to products in than 20,00 different grocery stores. As one of the primary leaders in grocery delivery, analyzing their datasets can be a real value in understanding how what customers are ordering and when.

Therefore, we propose to use the datasets provided on Kaggle by Instacart. https://www.kaggle.com/c/instacart-market-basket-analysis / . The datasets uses transactional data from customers that order through Instacart. The dataset was provided for Instacart's Market Basket Analysis competition, where the goal is to predict if a customer how likely and which products a customer is likely to re-order based on past purchases. We propose to attempt to make inferences on this based on our analysis.

The dataset contains order histories of around 200,000 customers. There are 7 tables total, one which is a training dataset for "orders. The others are related to the products, their aisle, past orders, and all orders of products. Below I have listed the 7 datasets, and the attributes within each of them.

Data Table Name:	Aisle.csv	
Attributes:	aisle_id (num)	Aisle (str)

Data Table Name:	departments.csv	
Attributes:	department_id (num)	department (str)

Data Table Name:	Order_product_prior.csv

Attributes:	order_id (num)	Product_id (num)	Add_to_cart_order (num)	Reordered (num)

Data Table Name:	Order_product_train.csv			
Attributes:	order_id (num)	Product_id (num)	Add_to_cart_order (num)	Reordered (num)

Data Table Name:	orders.csv							
Attributes:	order_id (num)	user_id (num)	Eval_set (str)	Order_number (num)	Order_dow (num)	Order_ hour_of_day (num)	Order_hour_of_day(num)	Days_since_prior_order (str)

Data Table Name:	products.csv			
Attributes:	product_id (num)	Product_name (str)	Aisle_id (num)	Department_id (num)

Data Table Name:	Aisle.csv	
Attributes:	order_id (num)	products (str)

Preprocessing:

Because the challenge of the Instacart dataset comes from the sheer volume data across the 7 different data tables, the dataset did not require much cleaning or preparation. Instead, the focus of preprocessing was to determine the scale of the datasets based on the unique ids and merging the various datasets to answer our questions. Therefore, pre-processing involved a high level of reordering, merging, pivot tables, indeces, identification and removal of duplicates or unique values, and aggregation. For example, we first utilized the orders_df dataset, to understand the unique number of users using Instacart and the number of products they bought. After, the size of our training and test data by the number of users was determined. To achieve this objective, we created a function to get the unique count of the “evaluation_set” of the “orders_df” data table. Our function aggregated the unique count of three evaluation types.

Additionally, for to answer many of our questions (for example, like about the popularity of products and user ordering habits) we had to merge two datasets together. We then created pivot tables, grouped by a variable (like dates or times) and then reordered the data. In the case of finding what days and times per week users ordered products, we also had to create a mean value of the number of products ordered during that time period on a certain day.

Two of them major preprocessing steps in the product analysis phase involved merging product details with the order_prior details in order to create a merged dataset on the top products ordered previously. This step led to the analysis of what the most popular aisle and departments were based on order history of users. The second major preprocessing step required using association methods to group product_id’s with the count of that item to determine how much of that specific product was ordered. These products were then plotted by sales. He resulting outputs helped understand the most popular products for re-ordering on Instacart.

Method of Analysis:

The analysis of the Instacart dataset involved using qualitative and quantitative methods to analyze answer the following proposed research questions:

- What are some of the most popular items?
- What are some of the most popular aisles/departments?
- What is the highest requested time of day or days for ordering?
- What is the lowest requested time of day or days for ordering?
- How many orders are new orders?
- How many orders are re-orders?
- How often to customers order?
- Do people re-order more for items in any certain department or aisle more than others .
- In general, how does re-ordering data compare to regular ordering of customers?

These methods included utilizing summary statistics, association, time of day/day of week features, and product description and user information. In order to achieve our results, our methods required analysis at the user-and-product levels, in addition to calculating a re-order ratio. Our output files are in the form of condensed data tables and graphs. Since the overall of this project is to predict which products a customer is likely to re-order based on past purchases, we focused our analysis on understanding the top products sold and re-ordered through Instacart.

This analysis is important for many levels of the grocery and food industry. First, at the grocery partners level, it allows for stores to better predict when key shopping hours are and what items are most bought by mobile shoppers. This can help stores better manage product placement and staffing. Additionally, it is important for supply chains and producers. For stores with high mobile shoppers there is a low chance of spontaneous shopping, which is last-minute buying that occurs once the patron is inside the store. This may force some producers to seek different partnerships with the grocery stores where they know a high volume of customers will pass by their product. Finally, this data is important for food analysts and industry chefs who are looking to better understand what consumers like to eat, in order to create or find the next food big trend.

Description of Program:

The program overall goal was to predict what product a user is likely to order next, based on past order data. The program was overall a two-leveled exploratory analysis where the first step was the analyze the user data and the second level looked for trends within the product datasets.

Conclusion:

The results of our analysis were the following:

On average, users had about 16 -17 orders, with about 10 products per order. However, the most common number of products ordered was 5. There were over half a million orders that had either 4, 5, or 6 products per order.

Highest user ordering was found to be on Sunday, followed by Monday. Overall we had a higher count of ordering in the beginning of the week. Therefore, we predict that orders are 30-40% more likely to occur on Sunday or Monday than another day of the week. In addition, regardless of the day, orders were mostly placed around the middle of the day. Ordering sees a heavy rise after 7am, until about 5pm. This is likely due to users placing orders while at work. Interesting that the largest number of users placing orders occurs on Monday around 12pm.

Additionally, when looking at user ordering behavior, we found that around half of customers re-ordered within about a week of their last order. However, there is a strong decline after 7 days, meaning less customers ordered between 7-29 days after their last order. Additionally, these users re-orderd the same products about 59% of the time.

With regard to what users were ordering, we found that most users ordered Produce (fresh fruits and vegetables) from Instacart. This was followed by Dairy/Eggs and then by snacks, beverages, and frozen foods. Users did not prefer to use Instacart to order Pet and bulk items. The most popular individual items bout by customers was Bananas. In fact, the second most popular product was actually organic bananas. This was followed by Organic Strawberries, and then Organic Baby Spinach.

Therefore, based on our analysis- we can expect that customers are likely to buy fresh fruits or vegetables 1-7 days after their last order. In this order, we can expect them to either buy organic or nonorganic bananas.