DS 3001: Foundations of Machine Learning

4 October 2024

Instacart Project: Data Wrangling/EDA Write-Up

**Data Overview and Key Variables:**

The data from this 2017 Kaggle “Instacart Market Basket Analysis” by Stanley et al. consists of three million grocery store orders from 200,000 users. For each user, there is anywhere between four and 100 orders documented. For each order, the sequence of products, hour of purchase, and day of the week is also documented. Lastly, the time between orders for each customer is recorded (Stanley et al., 2017).

The key variables of the data are initially split into five distinct dataframes, which are categorized by variables associated with overall Instacart orders, the products themselves, the aisles that the products come from, the departments associated with these products, and orders and products dataframe that combines information from the specific orders and products sets, as well as information related to reordering products (Stanley et al., 2017). We merged the two dataframes related to ordering into one dataset called complete\_orders\_df. Within this overall orders set, there are a few categories. Order\_id, user\_id, and product\_id represent unique order identifiers. The numerical values are order\_number, order\_hour\_of\_day, days\_since\_prior\_order, and add\_to\_cart\_order. They represent the position out of total orders per customer, time of day, the time between orders, and the chronological order in which a product was added to an order, respectively. Order\_dow and reordered are coded with numbers; for the day of week, 0 represents Saturday and 6 represents Friday, and for whether or not a product has been previously purchased, 0 corresponds to ‘no’ and 1 corresponds to ‘yes’ (Stanley et al., 2017).

In terms of creating an overall products dataset, we first merged the products and aisles data frame, followed by merging this result with the departments dataframe into a dataset called complete\_products\_df. Within this overall products dataset, the unique identifiers and numeric variables are product\_id, aisle\_id, and department\_id, while the categorical variables are product\_name, aisle, and department (Stanley et al., 2017).

**Usage of Data for Project Purpose and Expected Challenges:**

For our project, we are interested in understanding patterns of consumer behavior when it comes to grocery shopping. This Instacart data will give us a general understanding of grocery store habits. Understanding past purchases and decision making factors will allow us to make predictions about future purchasing, specifically to see if previously purchased products will be ordered again. We can also explore what days or times of day are most common when purchasing–even looking at the types of food purchased at, for example, late at night.

In pursuing this project, one challenge we expect to encounter is merging the different data sources. We are provided with information on grocery store aisles, departments, and products, and then we are given information on the customers. We will have to fully merge the datasets so that we can get a better understanding of what purchases were made. Additionally, we will have to combine situational order data (day of week, hour of day, days since prior order) with descriptions of what the order entails. One way we have already joined files is by linking the asile\_id and department\_id from the products csv with aisles csv and departments csv respectively to get more information. Another challenge we faced with the data is saving and running the new merged dataset in Colab. Combining products with orders creates a very large amount of data that is too big for github to handle. We are currently brainstorming ways to save our new data. One idea that we have is to drop certain variables in either of the overall products and orders datasets that potentially do not have very much relevance in the patterns and predictions we hope to explore with this project. This way, the full merged dataset will be smaller and hopefully will run successfully in Colab.

Once we are able to run the fully merged order and product datasets, we can further explore the relationships between what products consumers order. As an example, we can cross-tabulate between individual users based on their user\_id and the number of days\_since\_prior\_order to examine patterns in ordering, as well as between user\_id and reordered to examine patterns among individuals and the products they repeatedly purchase. We can also examine ordering patterns based on different days of the week or hours of the day, specifically if particular products are ordered more on particular days or times of day. These are several examples of relationships and patterns we intend to explore once we are able to load our fully merged orders and products datasets.

**Cleaning Variables:**

In cleaning both of the overall orders and products datasets, we first sought to identify any missingness and replace these values with ‘nans.’ For example, through examining the first five rows of the overall orders dataset, the variable days\_since\_prior order, we saw that there were a lot of “NaN” values. In attempting cleaning this variable, we coerced it to numeric and tried to replace any potential missing values with a nan value. Coercion of this variable to numeric and replacement of missing values with nans did not change the total count of the variable, however, and when the total number of missing values for this variable were calculated, the result was 2,078,068. This is due to the fact that missing values seem to already be classified by “NaN,” so there are no other blank or null values for this variable that could have been replaced.

Another variable that we cleaned was the initially numeric order\_dow variable from the overall orders dataset. The days Saturday through Friday, chronologically, were originally classified as 0 through 6. We added a new column called ‘order\_day’ to the overall orders dataset that associated and replaced each number with its corresponding date. We also cleaned the initially numeric reordered variable from the overall orders dataset, as an observation of 0 corresponded to the product not being reordered, while an observation of 1 corresponded to the product being reordered. We added a new column called ‘reordered\_yes/no’ to the end of the overall orders datasetthat associated and replaced the 0s and 1s with their corresponding ‘no’ and ‘yes.’

**Exploratory Data Analysis (EDA) - Plots and Descriptive Statistics Tables:**

We conducted some initial Exploratory Data Analysis (EDA) with the overall products and orders datasets to gain an understanding of their respective patterns. As we work to resolve the issue of being able to load the large, fully merged dataset, we will continue to conduct EDA to examine the relationships between products, order patterns of individual consumers, and ordering patterns overall, particularly with date and time-related data. The code for our EDA can be seen in the shared Github repo under the file “Project\_Phase\_1\_EDA\_Wrangling.ipynb.” For the overall products dataset, we examined the shape of the dataset, types of variables involved, and the column names to get an initial sense of the overall dataset. We examined the frequency of department distributions specifically, first using .unique() and .value\_counts() to observe the number of unique values as well as frequencies of department types, and visualized this using a histogram, from which we observed that “personal care,” “snacks,” and “pantry,” were the three highest occurring product department categories. We then examined aisles using .unique() and .value\_counts(), also visualizing this variable related to products with a histogram. From this histogram, we saw that the most commonly occurring aisle type was “missing,” followed by “candy chocolate,” “ice cream ice,” and “vitamins supplements.”

In the overall orders dataset, we initially examined its shape, the types of variables in it, and the column names initially. We looked at the .value\_counts() and .describe() results of the “order\_day” new column to see the most popular days and also visualized this through a histogram; Saturday and Sunday were the most popular days. We also made a kernel density plot of the original order\_dow variable, which produced the same results. In addition, we created a histogram plotting the order\_hour\_of\_day variable, with the most popular hours of the day being in between the 9 to 17 hour range. We also made a boxplot of the add\_to\_cart\_order variable, with the 50% quantile being seen around the value 6 to 7. Finally, we created a cross-tabulation between the reordered\_yes/no variable and the order\_day variable. From this cross-tabulation, we observed that Saturday and Sunday were the days that repeat orders were most likely to occur.

As previously mentioned, once we are able to load the fully merged dataset and perform EDA with this, we intend to explore relationships between variables associated with the overall products and orders datasets. In particular, we are interested in the relationships between particular users and the number of orders they make, the number of days in between their orders, the specific products they reorder, and how often they reorder products in general. In addition, we are interested in the relationships between specific products and the times of day that they are ordered, the days of the week that they are ordered, and their likelihood of being reordered. As we continue to develop this project, we will continue to brainstorm more relationships of interest.

Sources

Stanley, J., Risdal, M., Sharathrao, & Cukierski, W. (2017). Instacart market basket

analysis. *Kaggle*. <https://kaggle.com/competitions/instacart-market-basket-analysis>.