DS 3001: Foundations of Machine Learning

4 November 2024

Instacart Project: Pre-Analysis Plan

Our study is focused on customers as the main observation. We believe that this will provide the most interesting predictions and observable patterns. Customers are identified in our datasets by their unique "user_id" variable. The main question that will guide our inquiry is what key factors most strongly influence the purchasing behavior of Instacart users? Our supporting ideas include predicting what products individual customers will order repeatedly based on previous orders and when customers are most likely to place orders.

We plan to focus on supervised learning, which relies on identifiable datasets and predictor variables to produce an identifiable result and to optimize algorithm efficiency and correctness (Delua, 2021). We will use regression, which produces a numeric prediction result, in our predictions around the days of the week and times unique customers are most likely to order, as an example. On the other hand, classification, which produces a categorical prediction result, will be used in our categorical predictions, like which products unique customers are likely to repeatedly order. We also plan to make predictions using decision trees, especially if there is an element of non-linearity. It will be interesting to compare the measurable results, such as R-squared and mean squared error, from regression/classification to decision tree results. When we explore non-linear relationships, such as days_since_prior_order and reordered_yes/no, decision trees may capture the variance of predicted variables better than traditional regression or classification.

While we will focus on supervised learning, we plan to explore how PCA, an unsupervised learning algorithm, will impact a model's performance by reducing

multicollinearity and avoiding overfitting. We will compare instances of linear regression and classification models, evaluating their performance both with and without applying PCA before analysis. So, essentially, we'll run a regression on the original variables, then perform PCA and regress again on the resulting explanatory variables. The models will likely perform differently, given how multicollinearity can cause overfitting and skew predictor results. Multicollinearity may arise between highly correlated variables, such as order_day_of_week and order_hour_of_day.

For our analysis, many categorical variables need to be one-hot encoded, or transformed into numeric variables, to perform linear regression. Some of these variables include product_name, aisle, and department. Once these variables are transformed, we will be able to run the models. For regression, the two numeric variables that can be predicted are days_since_prior and add_to_cart_order. For the former, order_day_of_week, order_hour_of_day, or order_number may all influence why a customer's purchase sequence is the way it is. We can also run a regression of days_since_prior on the department to see if two orders are close together due to different grocery store categories. When predicting add_to_cart_order, product, department, and aisle will be the most relevant variables. For classification, we can perform regression on categorical variables, such as reordered_yes/no, order_day_of_week, aisle, and department. As an example, features like order_day_of_week, aisle, and days_since_prior_order might influence whether or not a customer reorders a product (reordered_yes/no). We will most likely encounter multicollinearity in both regression and classification and principal component analysis, or PCA, will be useful to tackle highly correlated variables. Additionally, we will use *K*-means clustering as our final unsupervised learning strategy, which applies well to an extensive dataset like ours. This can help us identify

patterns in customer behavior, identifying commonalities like grocery shopping day of week, hour of the day, or preferred department/aisle. We will select k, using the elbow point method, to be the centroids, apply maxmin normalization to the variables of choice, likely variables that we seek to identify relationships between such as user_id, order_day, days_since_prior_order, reordered_yes/no, and others, and conduct *K*-means clustering, performing multiple iterations.

We will know if our approach works when we can produce an identifiable pattern in our predictions. Success for our project will mean being able to make conclusions about which customers are most likely to order which products based on unique product identifiers and reordering data and when they are most likely to place these orders based on days of the week and times of day.

Success for regression and classification is determined by R^2 and RMSE. Success is most possible when we remove any outliers, ensure that the sum of squared error (SSE) is the smallest that it can be, and ensure that we are not underfitting or overfitting with our model. The closer the value is to 1, the better fit the model is. A negative R^2 indicates severe over-fitting. However, it is not always a poor output because it can still communicate something about the data. RSME is also a measure of fit that explains how close real values are to predicted values, with lower values indicating a better fit. PCA is a method used to combat collinearity, where the number of variables is reduced but information is maintained (IBM, 2023). We will define success with PCA as obtaining the least sum of squared error values. Success with *K*-means clustering will be observed in identifying relationships between variables that we did not previously hypothesize about. In terms of presenting our results, we could produce a table of regression coefficients after performing linear regression. PCA is often displayed in a scatter plot

and an arrow shows the relationships between values. High loading indicates a high correlation (IBM, 2023).

Our original weakness was regarding the size of the dataset. We have information on 3 million orders made by 200,000 users. The dataset is organized by individual products purchased in each order by each user, so there are over 32 million rows. As a result, we ran into size issues, both in merging the orders and products data sets and in being able to upload the data set back to GitHub. Since we are focusing on customers, we decided to order the data by customers who made the most orders. The range of possible total order numbers is 4-100. In order to satisfy the GitHub requirement, we limited the data to 2,900 users whose total number of orders ranged from 78-94. We were able to successfully merge this smaller dataset with all the information on products. The final dataframe contains 2,492,060 rows which is the number of total products purchased by those unique customers. Our final merged dataset is in parquet format in GitHub.

Due to the smaller subset of data we are choosing to work with, another weakness we anticipate being an issue is the accuracy of the model to predict customer behavior. The range of 74-99 total orders presents both positive and negative attributes. On one hand, we are looking at the customers with the biggest purchase habits, so it will be interesting to see how frequent spending influences behavior. On the other hand, we are eliminating the majority of possible total order numbers, so we will not have information on what a more normal or average purchasing behavior entails.

Lastly, while our data does contain millions of products, we have identified a lack of overlap between products. For example, kombucha may be a common item, however, there are many different flavors available. As a result, we plan to conduct most of our models using aisle

and department as they show groupings between products that are not dependent on the individual item.

We understand these weaknesses are the nature of the data, and will accept as best results from the model as possible. Regardless, we will learn about how extremely large datasets operate and walk away with skills to manage them.

References

Delua, J. (2021). Supervised versus unsupervised learning: what's the difference? *IBM*. https://www.ibm.com/think/topics/supervised-vs-unsupervised-learning.

IBM. (2023). What is principal component analysis (PCA)? *IBM*. https://www.ibm.com/topics/principal-component-analysis.