Executive summary

Thanos Kritikos
Haoran Li

The challenge of the current project is to predict the products from the users' upcoming order utilizing data from the Instacart app. Stakeholders in this work may include the company itself, as well as other online grocery shopping companies and app users. The current project's goal is to visualize Instacart data, categorize app users based on their customer behaviors, and predict their future purchases using machine learning algorithms.

The database was obtained through a kaggle.com competition. It includes over 3 million grocery orders from over 200,000 Instacart users. A relative measure of time between orders, prior orders, and a training data set are also provided, as well as the week and hour of day the order was placed. Data cleaning and merging occurred. The values that did not exist were replaced with 0.

Consumer trends were investigated using an exploratory analysis. The top purchased products were mostly fresh produce such as bananas, strawberries, spinach, and avocado. The product department has the most purchases, followed by dairy eggs, snacks, and so on. Monday and Tuesday are the most profitable days, and 10 a.m. to 4 p.m. is the most profitable time. People usually order around 5 items, and only 59% of all items have been reordered.

To begin, we want to categorize customers based on their shopping habits. We attempted principal component analysis by grouping customers into clusters. We used principal component analysis, which involves selecting the top six components and plotting them against each other before settling on one most significant pair. Based on that, we can divide customers into three distinct clusters using the K-means method. The clusters that we noticed were one more generic user profile, one with baby included (a lot of baby formula purchased) and one customer profile with a lot more fresh produce purchased.

The extreme gradient boosting model (XGBoost) is then used to train models that can be evaluated by generating scores on a few prominent features. In our case, the purchase amount of a specific customer and product, as well as its reorder rate, are the most important factors in predicting reordering patterns. We use grid search cross-validation, which is a common practice for adjusting parameters, to find the appropriate parameters for our modeling. We run XGBoost models on each cluster to get a more targeted model, which proved to be more efficient and accurate in predictions.

Overall, through this work the customer trends were analyzed. Customers' classification algorithms, as well as customer order prediction algorithms, were provided. The effects of the XGBoost features can be analyzed in the future to improve accuracy/F1 score, and other machine learning can be tested.