K-Means Clustering for Instacart Recommendations

Shuvechchha Kunwar, under the direction of Dr. Schrementi

Department of Mathematics and Statistics

Abstract

This paper is inspired by the extensive use of Recommendation Systems in this digital era. It draws concepts from Machine Learning and Data Science to develop a recommendation model employing Instacart's User Dataset. It aims to utilize the concept of collaborative filtering which predicts relevant products based on the behavior patterns of similar users. K-Means Clustering is used to split customers into distinct groups depending on their attributes. The predictions are made for each cluster of users based on the cluster's collective purchase pattern.
Keywords: recommendation, k means clustering, customer segmentation

Major Themes

The vast volume of data generated by businesses has piqued interest in the analytical world. The prediction model is one of the most prominent methods for gaining significant value from the collected data in a variety of ways, including analyzing customer behavior to create product recommendations, developing pricing models, and so on. This study will look into recommendation systems and their growing popularity in a variety of businesses, as evidenced by Netflix, Amazon, and Facebook, among others.

A **recommendation system** is a form of prediction model that predicts relevant products or services for consumers. For instance: Based on similarities such as genre, and ratings of the movies we've watched on Netflix, its recommendation system generates movie suggestions.

The most popular form of recommendation systems is content-based filtering and collaborative filtering. **Content-based filtering** focuses on the attributes of the product to recommend similar products to the users. **Collaborative filtering**, on the other hand, predicts relevant products or services for individual users based on the attributes of products used by users who have similar traits.

Customer segmentation is the method of categorizing consumers into groups based on shared traits, allowing firms to adapt their business models to meet the demands of their customers and personalize marketing strategies to better target each cluster of users. In this study, we focused on **K-Means Clustering** for our customer segmentation. K-Means Clustering is an unsupervised machine learning algorithm that groups customers into distinct, non-overlapping subgroups.

Dataset

The dataset employed in this study was part of Instacart's Kaggle Competition to improve its prediction model for recommending products to its users. This anonymized data file was made up of six relational files that contained information on the user's orders, such as product ids, names, aisles, and departments. It featured data on 3 Million grocery orders placed by 200,000 Instacart users. There were between 4 and 100 orders for each user, with the sequence of products purchased in each order. It also provided information on the day of the week and the hour of the day the order was placed.

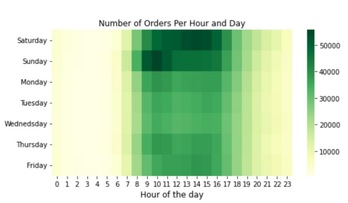
Approach and Tools

This research was divided into three main sections

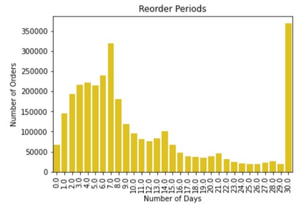
- **Exploratory Data Analysis** incorporated understanding the important attributes of data. Some of the analyses include determining the most often ordered products, aisles, and departments, as well as the most popular hour of the day and week of the day to place orders.
- **Customer segmentation** involved using K-Means Clustering to divide users into several clusters depending on their attributes. Because the dataset was huge, we randomly selected 25% of the users from the dataset for our analysis.
- **Recommendations** for each cluster were generated by summarizing the cluster's collective purchase pattern. This method of recommendation fit the concept of Collaborative Filtering, which was discussed previously.

Google Colab was used as the web IDE for Python. To learn about the dataset and perform operations on it, packages like numpy, pandas, sklearn, and seaborn were used.

Result 1



(a) Figure 1



(b) Figure 2

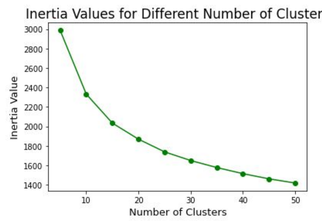
Figure 1 shows a heat map of the relationship between the number of orders and the hour of the day and day of the week. According to the plot, the majority of orders are placed between Saturday afternoons (1-3 PM) and Sunday mornings (9-10 AM).

Figure 2 depicts a bar plot on the relationship between the total number of days it takes users to reorder. According to the bar graph, most users reorder after a week (7 days) or a month (30 days). Smaller reorder peaks were also detected at 14, 21, and 28 days.

Acknowledgment

I would like to thank Dr. Schrementi for his unwavering guidance and encouragement throughout the semester. I would also like to thank the Mathematics and Statistics Department for providing this wonderful opportunity.

Result 2



(c) Figure 3

cluster	recommended_products
6	India Pale Ale
7	White Giant Paper Towel Rolls
8	Organic Strawberries
9	Half & Half
10	Organic Baby Spinach

(d) Figure 4

The K-Means Algorithm was used to cluster the data on the percentage of products ordered from each of 21 Departments of Grocery Stores by each randomly chosen user. Figure 3 depicts the elbow plot created to determine the best k-value for customer segmentation. According to the plot, 20 appears to be an optimal k-value for our model.

We divided the users into 20 clusters and found the most popular product in each cluster. Bananas were excluded from the data set during this analysis so that the most ordered product in each cluster would not be influenced. The most popular products for clusters 6 to 10 are shown in Figure 4.

Conclusion

When evaluated with train data from Instacart Data Files, this collaborative recommendation system correctly predicted a product that would be in a user's next order 9.82% of the time. When we use this algorithm on a broader scale, this accuracy in generating predictions might result in higher revenues for firms.

Limitations and Future Research

One of the key limitations of this study was the large size of the dataset, which impacted the memory usage of Colab and restricted us to using only 25% of the original dataset.

Future research could include more features to segment customers into distinct clusters. Similarly, an additional study on Product Attributes might be conducted in order to identify complementary products to each cluster.

References

[1] Chinedu Pascal Ezenkwu, Simeon Ozuomba, and Constance Kalu. Application of k-means algorithm for efficient customer segmentation: A strategy for targeted customer services.

[2] Phongsavanh Phorasim and Lasheng Yu. Movies recommendation system using collaborative filtering and k-means. *International Journal of Advanced Computer Research*, 7:52–59, 02 2017.

[3] Yini Zhang and Chenyun Zhu. Market basket analysis on instacart datasets. Unpublished Manuscript, 2017.

kunwars@hollins.edu