Leticia Lima Huang CIS9660- Data Mining for Business Analytics Technical report

This project centers on Customer segmentation analysis using a Kaggle dataset, reduced from 49,782 to 44,804 rows after cleaning missing values and negative quantities. This project comes with a notebook and a dashboard. Those tools are helpful for business to visualize past customer behavior and adapt the notebook to make predictions with their own dataset.

For the initial analysis, I used RFM to scores buyers based on Recency, Frequency and Monetary expenditures. Additionally, new columns were created to aid in visualizing different metrics such as average purchases, product diversity, seasonality, top products and return rate. These visualizations provided insights into common customer behaviors. Feature scaling was then conducted to prepare the data for unsupervised K-Means model. The optimal number of clusters identified was 2 with a Silhouette score of 32.3%. K-Means clustering produced two groups: Cluster 0 contained 35.2% of customers, while Cluster 1 included 64.8%. Recency and frequency metrics were similar between clusters; however, monetary expenditure differed significantly. Cluster 0 had an average spend of $2333.90, and Cluster 1 averaged $572.86. The average item price was $73.68 in Cluster 0 compared to $36.57 in Cluster 1. This indicates that Cluster 0 had higher spending and tended to purchase higher-priced items, while Cluster 1 spent less. Other features did not demonstrate significant differences.

Principal Component Analysis (PCA) was conducted to assess the explained variance within the dataset. The first principal component accounted for 50.5% of the variance, while the second explained 20.2%, resulting in a combined total of 70.7% of the dataset's variance captured by these two components. Cluster profile comparisons revealed minimal significance, as the clusters exhibited substantial similarity across most features. Hierarchical Clustering demonstrated a similar index of 0.646 with K-Means, although some differences were observed among subgroups. Additionally, the Apriori Algorithm was applied to categorize products into different baskets and identify product combinations frequently purchased together, utilizing minimum support and confidence thresholds. Association rules were then generated and ranked according to lift values to uncover potential relationships for recommendations or marketing initiatives. However, the resulting association rules displayed low values for support, confidence, and lift, indicating that product relationships are weak and found in a small subset of transactions. This suggests that enhanced marketing strategies are needed to increase customer regularity and efficiently promote products.

For cluster 0, our ROI projections estimate that a 25% response rate from re-engaging customers could result in an increase in revenue of $4,949,010. For cluster 1, which has lower engagement and consists of budget-conscious customers, the projected ROI is $2,241,851. This figure assumes efforts are focused on developing relationships and emphasizing promotional products and educational content to establish trust.