Customer Segmentation using Clustering Techniques

Prepared by: Abhishek Brahmbhatt

Date: 27-01-2025

This report outlines the results of customer segmentation performed using clustering techniques. The objective is to segment customers based on both their profile information (from Customers.csv) and transaction data (from Transactions.csv). K-Means clustering is applied to categorize customers into four distinct clusters.

The analysis also includes the evaluation of clustering metrics such as the Davies-Bouldin (DB) Index,
Silhouette Score, and Inertia to assess the quality of the clustering results.

Deliverables:

- Number of Clusters: 4
- DB Index: 1.391609754323979
- Silhouette Score: 0.2982928866134436
- Inertia: 3017.745613829656

This report provides a summary of the clustering methodology, the number of clusters formed, and key evaluation metrics.

Methodology

The segmentation process involved loading both customer profile data and transaction data.

After preprocessing and feature engineering, K-Means clustering was applied. We experimented with the number of clusters,

choosing 4 clusters as the optimal solution based on evaluation metrics. The clusters were analyzed using various evaluation techniques,

including the DB Index, Silhouette Score, and Inertia.

Clustering Evaluation Metrics:

1. DB Index: Measures the intra-cluster similarity and inter-cluster separation. A lower DB Index indicates better clustering.
2. Silhouette Score: Measures the cohesion and separation of the clusters. A score closer to 1 indicates well-separated clusters, while negative values suggest poor clustering.
3. Inertia: The sum of squared distances of samples to their closest cluster center. Lower inertia values indicate better clustering.

Business Insights

1. Number of Clusters: The K-Means clustering algorithm formed 4 distinct clusters based on customer profiles and transaction patterns.

2. DB Index: The DB Index value of 1.391609754323979 indicates a reasonably good clustering, with moderate separation between clusters.

3. Silhouette Score: The Silhouette Score of 0.2982928866134436 is low but still positive, suggesting that the clusters are moderately well-separated, but there may be room for improvement.

4. Inertia: The Inertia value of 3017.745613829656 indicates that the clusters are relatively compact, but further refinement could reduce the inertia and improve the quality of clustering.

5. Cluster Insights: Each cluster revealed different customer segments, such as high-spending

customers or those with specific preferences in product categories.

Explanation of Graphs

1. Cluster Visualization: A 2D scatter plot created using PCA shows the distribution of customers into 4 clusters.

The visualization helps understand the separation and composition of each group based on demographic and transactional data.

2. DB Index Plot: A plot representing the Davies-Bouldin Index is used to evaluate the clustering performance. A lower index indicates better separation between clusters.

3. Silhouette Score Plot: This graph visually represents the Silhouette Score for each sample, showing how well each customer fits within its assigned cluster.

Higher values indicate better cohesion within the cluster.

4. Inertia Plot: A graph displaying the Inertia over different iterations of clustering, highlighting the optimal number of clusters for minimizing within-cluster variance.

Conclusion

The customer segmentation process using K-Means clustering has resulted in four distinct clusters.

The evaluation metrics, including DB Index, Silhouette Score, and Inertia, provide insights into the clustering quality.

Future improvements can be made by exploring different clustering algorithms, adjusting the number of clusters,

or refining the feature engineering process for more accurate segmentation.

The following Jupyter Notebook/Python script containing the clustering code is provided for further reference and replication of the results.