

Customer Segmentation Report - K-Means Clustering

Objective:

The goal of this analysis is to segment customers based on their demographic and transactional data. The K-Means clustering algorithm was used to group the customers into clusters, enabling the identification of distinct customer segments for further analysis.

1. Number of Clusters Formed

- In this analysis, the K-Means algorithm was applied with **4 clusters** as the initial choice. The optimal number of clusters can vary, and experimenting with values ranging from **2 to 10** clusters is recommended for a more refined segmentation.

2. Clustering Evaluation:

DB Index (Davies-Bouldin Index):

The **Davies-Bouldin Index (DB Index)** measures the compactness and separation of clusters. A lower DB Index indicates better-defined clusters with distinct separation.

- **DB Index Value: 0.6022910408086608** (This value indicates that the clusters are relatively well-separated, but there's room for improvement in terms of compactness.)

Silhouette Score:

The **Silhouette Score** assesses how well-separated and internally consistent the clusters are. A higher score indicates that the clusters are both cohesive and well-separated.

- **Silhouette Score: 0.49625883499054974** (This score indicates that the clustering solution is fairly well-separated, though there may still be some overlap between clusters. A value closer to 1 would indicate a more distinct separation.)

3. Visualizations:

To visually inspect the clustering results, Principal Component Analysis (PCA) was performed to reduce the dimensionality of the data to 2D. This allows us to plot the data points in a 2D scatter plot, where each point is colored based on its cluster label.

The plot below shows the customer segmentation in 2D space:

Interpretation of Visuals:

- The scatter plot demonstrates the separation of the customer segments based on the features used for clustering.
- Each color represents a distinct cluster, and the distance between points indicates how far apart the clusters are in the 2D reduced space.
- The clusters appear to be relatively well-separated, which aligns with the moderate Silhouette Score of 0.49625883499054974.

Conclusion:

- The clustering analysis reveals that the K-Means algorithm was successful in segmenting the customers into 4 clusters.
- The DB Index and Silhouette Score suggest that the clusters are reasonably well-separated and cohesive, though further adjustments (e.g., tuning the number of clusters or features used) could improve the clustering quality.
- These segments can now be analyzed further for targeted marketing or personalized recommendations.