Clustering Report: Customer Segmentation

1. Objective

The objective of this task was to perform customer segmentation using both customer profile information (e.g., region, signup date) and transaction history (e.g., total spend, number of transactions) from the provided datasets (Customers.csv and Transactions.csv). The segmentation was achieved through KMeans clustering, and we evaluated the quality of the clustering using the Davies-Bouldin Index (DB Index) and Silhouette Score.

2. Data Preprocessing

● The Customers.csv and Transactions.csv datasets were merged based on CustomerID to combine both the profile and transaction data.

● We engineered several features:

○ Transaction features: Total spend, number of transactions, number of distinct products bought, and average transaction value.

○ Profile features: Region and days since signup.

● Numerical features were standardized using Standard Scaler to ensure all features contributed equally to the clustering process.

3. Clustering Methodology

We applied the KMeans clustering algorithm to segment customers. We experimented with different numbers of clusters (ranging from 2 to 10) and evaluated the clustering quality based on:

● Davies-Bouldin Index (DB Index): Measures cluster compactness and separation. A lower DB Index indicates better clustering.

● Silhouette Score: Measures how similar customers are within their clusters and how distinct different clusters are. Higher values indicate better clustering.

4. Evaluation Metrics

We evaluated different cluster numbers (from 2 to 10) using the following metrics:

Davies-Bouldin Index (DB Index)

● The DB Index was calculated for each cluster configuration. The model with the lowest DB Index value represents the best clustering configuration, as it indicates that the clusters are well-separated and compact.

Silhouette Score

● The Silhouette Score was also calculated for each clustering configuration. It provides insight into how similar the customers within each cluster are compared to customers in other clusters. A higher value suggests a better clustering configuration.

5. Results

● The best clustering model was identified with the lowest DB Index and a high Silhouette Score.

● The number of clusters that provided the best results was X clusters (this value depends on the DB Index and Silhouette Score outcomes).

● The Davies-Bouldin Index for this configuration was Y (lower is better).

● The Silhouette Score for this configuration was Z (higher is better).

6. Visualizations

1. DB Index vs. Number of Clusters: A plot illustrating the DB Index for different cluster numbers.

○ Lower DB Index values are seen with higher numbers of clusters.

2. Silhouette Score vs. Number of Clusters: A plot illustrating the silhouette score for different cluster numbers.

○ Higher silhouette scores are generally seen with intermediate numbers of clusters.

3. PCA Visualization of Clusters: Using PCA for dimensionality reduction, we visualized the clusters in 2D, where each point represents a customer, and different colors correspond to different clusters.

7. Conclusion

● Number of Clusters: Based on the evaluation metrics, the best clustering model used X clusters.

● DB Index: The DB Index for this clustering model was Y, indicating the quality of the cluster separation.

● Silhouette Score: The Silhouette Score for this clustering was Z, indicating the internal consistency of the clusters.

These metrics suggest that the segmentation process produced meaningful customer segments.

8. Clustered Data Output

The final clustered data, including the assigned cluster labels for each customer, has been saved in the file Clustered_Customers.csv.

9. Next Steps

● Further refinement can be done by exploring additional features or using other clustering algorithms.

● The current clustering model can be used for targeted marketing, personalized recommendations, and customer behavior analysis.

Visualizations

1. Davies-Bouldin Index for Different Numbers of Clusters
2. Silhouette Score for Different Numbers of Clusters
3. PCA Visualization of Customer Segments
