Customer Segmentation Clustering Report

Overview

This report provides an in-depth analysis of customer segmentation using KMeans clustering, applied to a dataset that includes transactional and customer profile data. The goal is to segment customers based on their purchasing behaviour and tenure. Various clustering evaluation metrics, including the Davies-Bouldin Index (DB Index) and Silhouette Score, have been calculated to assess the clustering performance. The report also includes a visual representation of the clusters using Principal Component Analysis (PCA).

Data Preprocessing and Feature Engineering

The dataset includes customer and transactional data, which was preprocessed to extract relevant features for clustering. First, the transactions were aggregated by customer, resulting in metrics like the number of transactions, total spending, and total quantity purchased. In addition, customer tenure was calculated as the difference between the first and last purchase date, and signup tenure was calculated by measuring the time from account creation to the present.

We also handled categorical variables like "Region" by encoding them into dummy variables. The final features used for clustering included:

- **Transaction Count:** Number of transactions per customer.
- **Total Spend:** Total spending by the customer.
- **Total Quantity:** Total quantity purchased by the customer.
- **Customer Tenure:** Duration between the first and last purchase.
- **Signup Tenure:** Duration from account creation to the present day.

These features were then standardized using the StandardScaler to ensure that each feature contributed equally to the clustering process.

Determining the Optimal Number of Clusters

To determine the optimal number of clusters for the KMeans algorithm, the Elbow Method was employed. This method involves plotting the inertia (sum of squared distances from each point to its assigned cluster centre) for different numbers of clusters. The "elbow" point on the plot indicates the optimal number of clusters. In this case, the optimal number was determined to be **4**, as the inertia began to level off beyond this point, suggesting that increasing the number of clusters further would not substantially improve the model's performance.

Clustering Using KMeans

With the optimal number of clusters determined, KMeans clustering was applied with **4 clusters**. Each customer was assigned to one of these clusters, and the cluster assignments were added to the dataset for further analysis.

Evaluation of Clustering Quality

Two commonly used metrics for evaluating clustering performance are the Davies-Bouldin Index and the Silhouette Score.

- **Davies-Bouldin Index (DB Index):** This index measures the average similarity ratio of each cluster with the cluster that is most similar to it. A lower value of the DB Index indicates better clustering quality. For the current model, the DB Index was calculated to be **1.1057676220191057.** A lower DB Index suggests that the clusters are well-separated and compact.
- **Silhouette Score:** This metric assesses how similar an object is to its own cluster compared to other clusters. A higher silhouette score (closer to 1) indicates that the clusters are well-defined and separated, while a score closer to 0 indicates that the clusters are overlapping. The Silhouette Score for clustering **0.2978627659907817**, which indicates that the clusters are reasonably distinct.

Visualizing the Clusters

The clusters were visualized using PCA (Principal Component Analysis), a technique that reduces the dimensionality of the data while retaining its variance. A scatter plot of the two principal components was generated, with points coloured by their cluster assignment. This visual representation helps confirm the separability of the clusters, with each cluster occupying a distinct region in the 2D space.

Conclusion

The KMeans clustering model successfully segmented the customer base into **4 clusters**, with good performance as indicated by the **Davies-Bouldin Index** and **Silhouette Score**. These results suggest that the clusters are both well-separated and compact, making them useful for targeted marketing or customer behaviour analysis.

In conclusion, the segmentation model provides a solid foundation for understanding customer behaviours, and the next steps would be to apply the results in a practical context, such as personalized marketing or retention strategies.