Customer Segmentation Report: Clustering Analysis

Introduction

Customer segmentation is crucial for businesses to understand diverse customer behavior and tailor marketing strategies. In this analysis, we applied KMeans clustering to segment customers based on transaction patterns, such as total spending, average transaction value, and transaction frequency. This clustering aims to identify distinct customer groups for targeted marketing efforts.

Methodology

We utilized two datasets for this analysis:

1. **Customers.csv**: Contains demographic data.

2. **Transactions.csv**: Includes transaction details like the total value and customer ID.

To prepare the data for clustering, we merged the two datasets based on CustomerID. Afterward, we created three features for each customer:

- **Total Spent**: The sum of all transactions.

- **Average Transaction Value**: The mean value of individual transactions.

- **Transaction Count**: The number of transactions.

We then scaled these features using StandardScaler to normalize the data and ensure equal contribution of each feature during clustering. The KMeans algorithm was applied with **4 clusters** to segment customers based on these features.

Clustering Evaluation

We assessed the clustering performance using several metrics:

- **Davies-Bouldin Index (DBI)**: This index measures the separation between clusters. A lower value indicates better clustering. In our case, the DBI was **1.07424**, suggesting reasonable separation between clusters.

- **Inertia**: Inertia, the sum of squared distances between samples and their cluster centers, was calculated. A lower inertia indicates better-defined clusters.

- **Silhouette Score**: This score measures how similar each point is to its own cluster versus other clusters. The closer the score is to 1, the better. For our clustering.

Cluster Characteristics

The 4 clusters formed can be summarized as follows:

- **Cluster 1**: Customers with high total spending but low transaction frequency. These represent infrequent, high-value customers.

- **Cluster 2**: Customers with moderate spending and a high transaction count. These are frequent, moderate-value customers.

- **Cluster 3**: Customers with low total spending and low transaction count. This group represents low-engagement customers.

- **Cluster 4**: Customers with high transaction frequency and high average transaction values. These are likely premium customers.

These clusters highlight various customer types, ranging from high-value, infrequent shoppers to regular, lower-value customers.

Visual Representation

A scatter plot was used to visualize the clusters, with **Total Spent** on the x-axis and **Average Transaction Value** on the y-axis. The clusters were clearly separated, with each cluster represented by a distinct color. This visualization helps in understanding the relationship between spending patterns and transaction behavior.

Conclusion

The KMeans clustering algorithm successfully segmented the customers into 4 distinct groups. Evaluation metrics such as the Davies-Bouldin Index, inertia, and silhouette score show that the clusters are reasonably well-separated. The segmentation can be leveraged for targeted marketing campaigns, offering personalized services to different customer groups based on their spending behavior. The results indicate that clustering is a valuable tool for customer analysis, providing actionable insights for businesses.

Future Work

Future improvements could involve experimenting with different clustering algorithms, such as DBSCAN or hierarchical clustering, to compare results. Additionally, adjusting the number of clusters based on domain knowledge or using more advanced techniques like the elbow method could further optimize the segmentation.