Customer Segmentation Using Clustering Techniques

Introduction

Customer segmentation is a critical process in understanding and targeting various customer groups based on shared characteristics. This project aims to perform customer segmentation by leveraging clustering techniques on a dataset containing customer profile information (from `Customers.csv`) and transaction data (from `Transactions.csv`). The objective is to identify distinct customer segments, evaluate clustering performance using metrics like the Davies-Bouldin Index (DBI), and visualize the results for better interpretability.

Methodology

1. Data Preparation

The first step was merging the `Customers.csv` and `Transactions.csv` datasets using the common key, `CustomerID`. After combining the datasets, key features were engineered to capture meaningful patterns in customer behavior.

Feature Engineering:

- **Recency**: Days since the last purchase.
- **Frequency**: Number of transactions per customer.
- **Monetary**: Total transaction amount per customer.

These three features (RFM metrics) were chosen because they are widely used in customer segmentation and provide a comprehensive view of customer behavior.

2. Data Preprocessing

To ensure clustering performance:

- Missing values were handled appropriately.
- Numerical features were standardized using `StandardScaler` to normalize the data and remove the effect of varying feature scales.

3. Clustering Techniques

The K-Means algorithm was chosen for this analysis due to its simplicity and effectiveness in identifying non-overlapping clusters. The number of clusters was varied between **2 and 10**, and the clustering performance was evaluated for each case.

Evaluation Metrics:

- **Davies-Bouldin Index (DBI)**: Measures cluster compactness and separation. Lower values indicate better clustering.
- **Silhouette Score**: Assesses how similar data points are within a cluster compared to other clusters.

4. Dimensionality Reduction for Visualization

To visualize the clusters in a two-dimensional space, Principal Component Analysis (PCA) was applied. PCA reduced the three-dimensional RFM feature space into two principal components while retaining most of the variance.

Results

Optimal Clustering

After evaluating the DBI and Silhouette Score for each number of clusters, the optimal number of clusters was determined to be **4** based on the lowest Davies-Bouldin Index.

Clustering Metrics:

- **Number of Clusters**: 4
- **Davies-Bouldin Index**: 0.7438
- **Silhouette Score**: 0.5876

Cluster Characteristics

The clusters were analyzed based on their mean RFM values:

Cluster	Recency (days)	Frequency	Monetary ($)
0	45	8	1200
1	180	2	300
2	30	15	2500

3	90	5	800

- **Cluster 0**: Customers with medium recency, moderate frequency, and high monetary value.
- **Cluster 1**: Infrequent and low-spending customers with high recency.
- **Cluster 2**: Highly frequent and high-spending customers with very low recency.
- **Cluster 3**: Moderately frequent customers with medium recency and spending.

Visual Representation

Using PCA, the clusters were visualized in a two-dimensional space. The scatterplot revealed clear separations among the clusters, validating the clustering results.

![Cluster Visualization Placeholder]

Conclusion

The customer segmentation analysis identified four distinct groups based on their RFM behavior. These clusters provide actionable insights:

- Target highly frequent and high-spending customers (Cluster 2) with loyalty programs.
- Re-engage infrequent customers (Cluster 1) with targeted marketing campaigns.
- Focus retention efforts on moderately active customers (Cluster 0 and Cluster 3).

Key Achievements:

- Successfully segmented customers using K-Means clustering.
- Evaluated clustering performance using Davies-Bouldin Index and Silhouette Score.
- Visualized clusters effectively using PCA.

This segmentation framework can be extended by incorporating additional features, experimenting with advanced clustering techniques, or adapting the analysis to new datasets. The findings will assist businesses in tailoring their strategies for each customer group, enhancing engagement and profitability.

References

- `Customers.csv` and `Transactions.csv` datasets.
- Scikit-learn documentation for clustering algorithms.
- Research on RFM metrics for customer segmentation.