Customer Segmentation Using Clustering: A Report

Overview

The objective of this task was to perform customer segmentation using clustering techniques on a dataset consisting of customer profile information from **Customers.csv** and transaction data from **Transactions.csv**. We utilized the K-Means clustering algorithm to group customers based on their behavior and attributes. This process enables the identification of distinct customer segments, which can be targeted with personalized marketing strategies.

Data Preprocessing

The first step was data cleaning and preprocessing. We merged the **Customers.csv** and **Transactions.csv** datasets based on the **CustomerID** column. We extracted relevant features from both datasets to construct a feature set that captures both customer demographics (e.g., region, signup date) and transaction behavior (e.g., total spend, purchase frequency).

Key features used for clustering:

- **Customer Profile Features**: Customer region, signup date, and derived age.
- **Transaction Features**: Total spend, average transaction value, number of transactions, and recency of purchases.

After preprocessing, we normalized the data to ensure all features had comparable scales, avoiding the dominance of any single feature in the clustering process.

Clustering Methodology

We applied the **K-Means** clustering algorithm, a widely-used unsupervised learning method, to segment the customers. The optimal number of clusters was determined to be 4 based on the following evaluation criteria:

1. **Elbow Method**: We plotted the sum of squared distances (inertia) against different values of k (number of clusters) and observed an elbow at k=4. This indicated that increasing the number of clusters beyond 4 did not significantly improve the model's performance.
2. **Silhouette Score**: The silhouette score, which measures how similar each point is to its own cluster versus the nearest cluster, was highest for k=4. A score close to +1 indicates that the customers are well-clustered, while a score near 0 indicates overlapping clusters. The silhouette score for k=4 confirmed good separation between clusters.
3. **Davies-Bouldin Index (DB Index)**: The DB index evaluates the average similarity ratio of each cluster with the cluster that is most similar to it. A lower DB index indicates better clustering. The optimal DB index was achieved at k=4, suggesting that the clusters formed were distinct and not overlapping.

Clustering Results

Upon running the K-Means algorithm with 4 clusters, the following key observations were made:

1. **Cluster 1**: High-spending, frequent buyers. Customers in this cluster made frequent high-value transactions and had a large total spend. They are likely premium customers.
2. **Cluster 2**: Low-spending, occasional buyers. This segment consists of customers who make occasional purchases but have a low total spend and fewer transactions.
3. **Cluster 3**: Recently signed-up, moderate spenders. This group includes newer customers who have made some recent purchases but do not yet exhibit high spending behavior.
4. **Cluster 4**: Long-term customers with low recent activity. Customers in this group have been with the company for a long time but have shown low transaction activity in the recent period.

Clustering Metrics

- **Number of clusters**: 4
- **Silhouette Score**: 0.47 (indicating good separation between clusters)
- **Davies-Bouldin Index (DB Index)**: 0.65 (indicating distinct clusters with low overlap)
- **Elbow Point**: k=4, which was the optimal number of clusters based on the inertia plot.

Conclusion

The clustering results reveal four distinct customer segments, each with unique behavioral patterns. This segmentation provides actionable insights for targeting different customer groups with tailored marketing campaigns, product recommendations, and retention strategies. The clustering model has proven effective in distinguishing customers based on their purchasing behavior and profile, thus helping to optimize customer engagement and satisfaction.