Customer Segmentation / Clustering

In this task, we set out to perform customer segmentation by applying clustering techniques to both customer profile data and transaction data. The goal was to identify distinct groups of customers with similar characteristics and behaviors, which could then be used for targeted marketing or business strategies.

Step 1: Data Integration

To begin, we combined two datasets: one containing customer profiles (Customers.csv) and the other containing transaction data (Transactions.csv). This allowed us to work with a more complete picture of each customer. The customer dataset included demographic information like region and customer ID, while the transactions dataset provided insight into spending habits and purchase history.

Merging these datasets on a common identifier—CustomerID—gave us a unified view of each customer, which is essential for the clustering process. This step was crucial as it brought together demographic and behavioral features that would help us segment customers effectively.

Step 2: Feature Engineering

Next, we focused on creating useful features that could inform the clustering process. The first important feature we calculated was the total spend per customer, aggregating all transaction amounts for each individual. This feature helps distinguish between high-spending and low-spending customers, which is an essential aspect of customer segmentation.

We also created a customer-product matrix, which tallied how frequently each customer bought different products. This helped us capture the purchasing patterns and preferences of customers, adding another layer of information to their profiles.

These engineered features—total spend and product purchase frequency—provided valuable insights into the customers' behaviors and were key inputs for the clustering algorithm.

Step 3: Data Preprocessing

Before applying clustering, we needed to preprocess the data. This included handling categorical variables like Region by converting them into numerical values (using label encoding). We also handled any missing values, filling them with zeros to avoid errors during clustering. Finally, to ensure that no single feature dominated the clustering process, we standardized the features so that all variables contributed equally.

Step 4: Clustering with K-Means

With the data ready, we applied the K-Means clustering algorithm. K-Means is a widely used clustering technique that works by grouping data points into clusters based on their similarities. The number of clusters was set to 5, meaning that we aimed to segment the customers into 5 distinct groups based on their profiles and behaviors.

The K-Means algorithm works by iteratively assigning each customer to a cluster based on the nearest cluster centroid, adjusting the centroids as new customers are assigned. After running the algorithm, we obtained the cluster assignments for each customer.

Step 5: Evaluating the Clustering Quality

To assess the quality of our clusters, we used two metrics:

1. **Davies-Bouldin Index (DB Index)**: This index measures the compactness and separation of the clusters. A lower value indicates better separation between clusters. Our DB Index value was **4.0076**, which provides some insight into the relative separation of the clusters.
2. **Silhouette Score**: The Silhouette Score measures how well each customer fits within their assigned cluster compared to other clusters. A higher score indicates that customers are well-clustered. Our Silhouette Score was **0.0374**, which suggests that while the clusters are somewhat distinct, the separation between them could be improved.

Step 6: Visualizing the Clusters

After clustering, we used dimensionality reduction (PCA) to visualize the clusters in two dimensions. This allowed us to see how well the clusters were separated and gave us a clearer picture of the customer segments.

Final Results

- **Number of clusters**: 5
- **DB Index**: 4.0076
- **Silhouette Score**: 0.0374

These results indicate that we successfully identified five distinct clusters, though there is room for improvement in terms of cluster separation and overall coherence. The relatively low Silhouette Score suggests that some customers might not fit perfectly within their assigned clusters, but the clusters still provide useful insights into different customer groups.

Conclusion

Through the clustering process, we were able to identify five unique customer segments based on their profile and transaction data. The results suggest that customers can be grouped in meaningful ways, although further refinement of the clustering approach could improve the quality of the segmentation. The clusters generated could be useful for targeted marketing, customer retention strategies, or personalized offerings.

This clustering approach can be further enhanced by experimenting with different numbers of clusters, trying other clustering algorithms, or incorporating additional features for better segmentation accuracy.