Customer Segmentation / Clustering

- Jagruti Piprade

Indian Institute of Information Technology, Vadodara
Computer Science Engineering, Final year

The goal of this customer segmentation task is to categorize customers into distinct groups based on their profile and transaction history. Clustering is an unsupervised machine learning technique that helps in grouping similar data points together. In this case, the clustering is done using both customer demographic information (such as region and signup date) and transaction information (such as total spending, frequency of transactions, and product preferences).

We used the **KMeans** clustering algorithm to perform this segmentation and evaluated the performance of the clustering using the **Davies-Bouldin Index** (DB Index).

Clustering Process

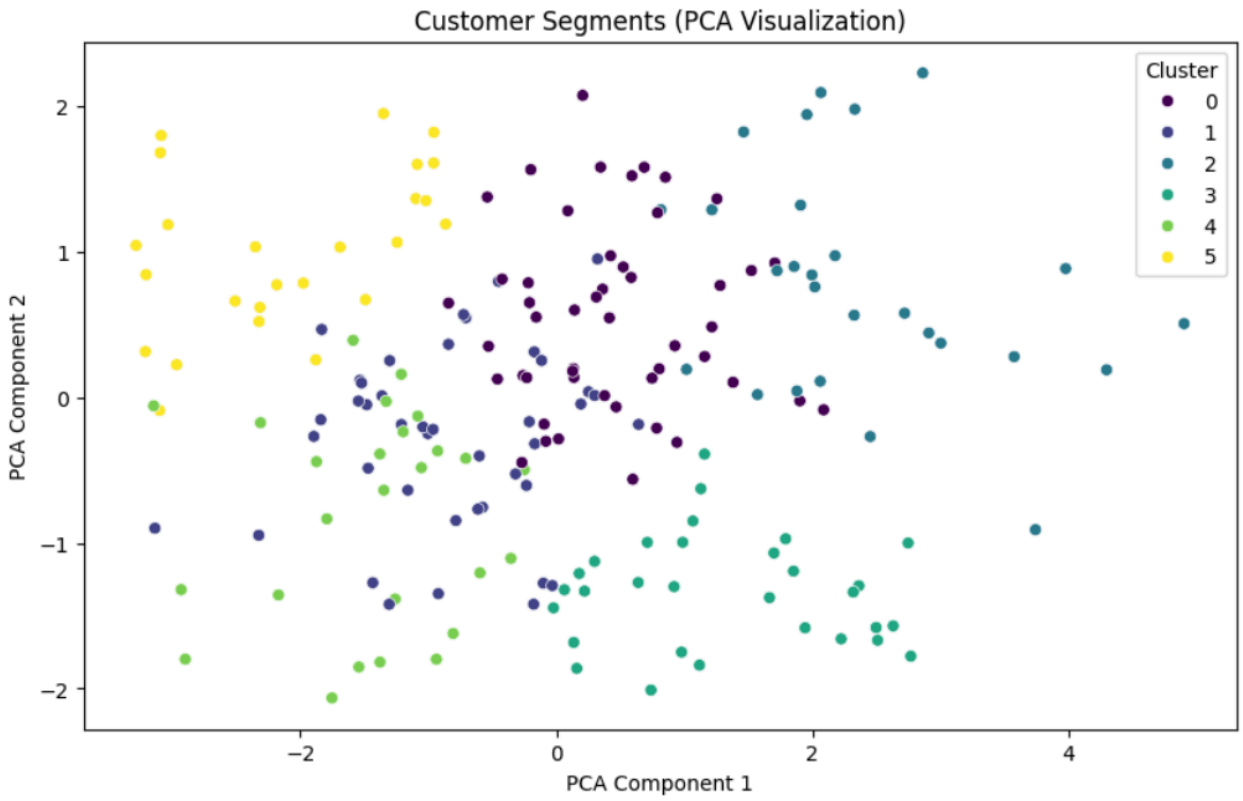
1. Data Preprocessing: The customer profile data from Customers.csv and transaction data from Transactions.csv were combined to create a feature matrix. This matrix was scaled for better performance during clustering.

2. Clustering Algorithm: KMeans clustering was chosen for this task. KMeans is a centroid-based clustering algorithm that partitions data into k clusters by minimizing the variance within each cluster. We tested various values of k (number of clusters) ranging from 2 to 10 to find the optimal number of clusters.

3. Optimal Number of Clusters: The number of clusters that minimizes the Davies-Bouldin Index was selected as the optimal k. The Davies-Bouldin Index is a metric used to evaluate the clustering results. A lower DB Index indicates better clustering as it signifies that the clusters are more distinct and less dispersed.

Results

- **Optimal Number of Clusters**: After evaluating clustering results for k values between 2 and 10, the optimal number of clusters was found to be 6.
- **DB Index**: The DB Index for the optimal clustering solution with k=6 was 1.1265. This value indicates that the clusters are well-separated and not overly dispersed.



Conclusion

1. **Optimal Number of Clusters**: 6 clusters were identified as the optimal number based on the lowest DB Index (1.1265).
2. **DB Index**: A value of 1.1265 suggests that the clustering model has formed well-separated clusters with reasonable compactness.
3. **Clustering Insights**: The segmentation of customers into 6 groups enables better understanding and targeted marketing strategies. By analyzing the characteristics of each cluster, businesses can design personalized offers, improve customer experience, and optimize resources.