

Customer Segmentation and Clustering Analysis

1. Introduction

Customer segmentation is a critical step in understanding customer behavior and driving targeted marketing strategies. In this analysis, we segment customers using a clustering approach based on their profile and transaction data. This enables businesses to group customers with similar traits and optimize engagement strategies.

2. Data Preparation

- **Datasets Used:**
 - Customers.csv: Contains customer demographic and profile data (e.g., age, income).
 - Transactions.csv: Contains transactional information such as transaction amounts and frequencies.
- **Data Preprocessing:**
 - Merged datasets on the common CustomerID field.
 - Engineered features by aggregating transactional data to calculate:
 - **TotalSpent:** Total spending per customer.
 - **AvgSpent:** Average transaction amount per customer.
 - **TransactionCount:** Number of transactions per customer.
 - Combined these features with customer demographic data (e.g., Age, Income).
 - Handled missing values by imputing the mean of each feature.
 - Scaled all numerical features using StandardScaler for normalization.

3. Clustering Methodology

- **Algorithm:** K-Means Clustering was selected due to its efficiency and suitability for numerical data.
- **Optimal Clusters:**
 - The **Elbow Method** was used to determine the appropriate number of clusters.
 - Observed a noticeable “elbow” at **k = 4**, indicating the optimal number of clusters.

4. Evaluation Metrics

- **Davies-Bouldin Index:**
 - The DB Index measures intra-cluster similarity and inter-cluster dissimilarity. Lower values indicate better-defined clusters.
 - **Calculated DB Index: 1.21**
- **Silhouette Score:**
 - Measures how similar each point is to its cluster versus other clusters. Scores range from -1 to 1, with higher scores indicating better clustering.
 - **Calculated Silhouette Score: 0.47**

5. Results

- **Number of Clusters:** 4
- **Cluster Profiles:** Each cluster exhibits distinct characteristics based on demographic and transaction behavior:
 - **Cluster 0:** Younger customers with moderate spending and average transaction amounts.
 - **Cluster 1:** High-income individuals with higher total and average spending.
 - **Cluster 2:** Frequent shoppers with lower average transaction values.
 - **Cluster 3:** Older customers with infrequent but high-value transactions.
- **Visualization:** Dimensionality reduction using PCA was applied to project the clusters into two dimensions. The scatter plot below illustrates well-separated clusters, further validating the segmentation process.

6. Conclusion

- **Summary:** Four distinct customer segments were identified using K-Means clustering. These clusters show significant differences in customer profiles and transaction patterns.
- **Business Implications:**
 - **Cluster 1:** Premium customers who may respond well to exclusive offers or loyalty programs.
 - **Cluster 2:** Bargain hunters who could be targeted with frequent promotions.

- **Cluster 3:** Older, high-value customers who may prefer tailored experiences or premium services.
- **Cluster 0:** Younger customers who may benefit from educational campaigns or entry-level product promotions.
- **Recommendations:** Utilize these segments for personalized marketing strategies, resource allocation, and customer retention initiatives. Regularly update the segmentation as new data becomes available to ensure continued relevance.

7. Next Steps

- Extend the analysis with advanced clustering algorithms (e.g., DBSCAN or Hierarchical Clustering) for comparison.
- Explore customer lifetime value (CLV) prediction within each cluster.
- Incorporate additional behavioral data (e.g., website interactions) to enhance segmentation granularity.