Task 3: Customer Segmentation / Clustering Report

Om Biradar

January 28, 2025

Contents

Task Description

The objective of this task is to perform customer segmentation using clustering techniques, leveraging both profile information (*Customers.csv*) and transaction data (*Transactions.csv*). The deliverables include:

- Determination of the optimal number of clusters (between 2 and 10).
- Calculation of clustering metrics, including the Davies-Bouldin (DB) Index.
- Visualization of the clusters using relevant plots.
- A detailed report containing clustering results and insights.

1 Methodology

1.1 Data Preparation

The data from *Customers.csv*, *Transactions.csv*, and *Products.csv* was merged to form a comprehensive dataset. The datasets were joined based on common keys (e.g., CustomerID, ProductID). Key steps include:

- Parsing date columns (e.g., *SignupDate*, *TransactionDate*).
- Combining transaction, customer, and product data for a holistic view.

1.2 Feature Engineering

To capture customer behavior, the following features were created:

- **RFM Features:** Recency, Frequency, and Monetary value.
- **Category Preferences:** Proportions of transactions in different product categories.
- **Purchase Patterns:** Average and standard deviation of purchase quantity, purchase span, etc.
- **Demographics:** Regional data and tenure (days since signup).

Sample Features Table:

CustomerID	Recency	Frequency	Monetary	Region
101	12	5	1500	North
102	30	2	700	South

Table 1: Sample of engineered features.

1.3 Data Preprocessing

The data was preprocessed using a pipeline:

- Numerical features (e.g., Recency, Frequency, Monetary) were standardized.
- Categorical features (e.g., Region) were one-hot encoded.
- Proportional features (e.g., category preferences) were passed through unchanged.

2 Clustering Process

2.1 Evaluation of Optimal Clusters

Clustering was performed using the *KMeans* algorithm. Metrics were calculated for clusters ranging from 2 to 10:

- **Davies-Bouldin Index:** Measures the quality of clustering (lower is better).
- **Silhouette Score:** Measures how well clusters are separated (higher is better).

Evaluation Metrics Plot:

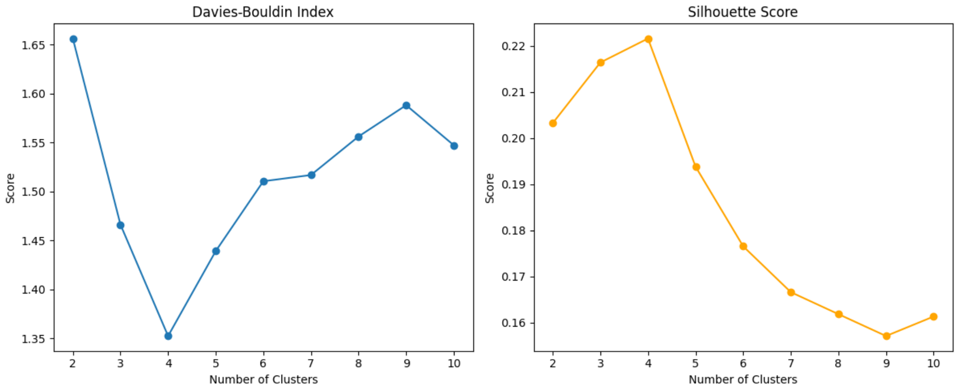

Figure 1: Davies-Bouldin Index and Silhouette Score for different cluster numbers.

2.2 Final Clustering

The optimal number of clusters was determined to be **4**, based on the evaluation metrics. Clustering was performed, and cluster labels were assigned to each customer.

3 Visualizations

3.1 PCA Visualization

Principal Component Analysis (PCA) was used to reduce dimensionality for visualization. The resulting clusters are shown below:

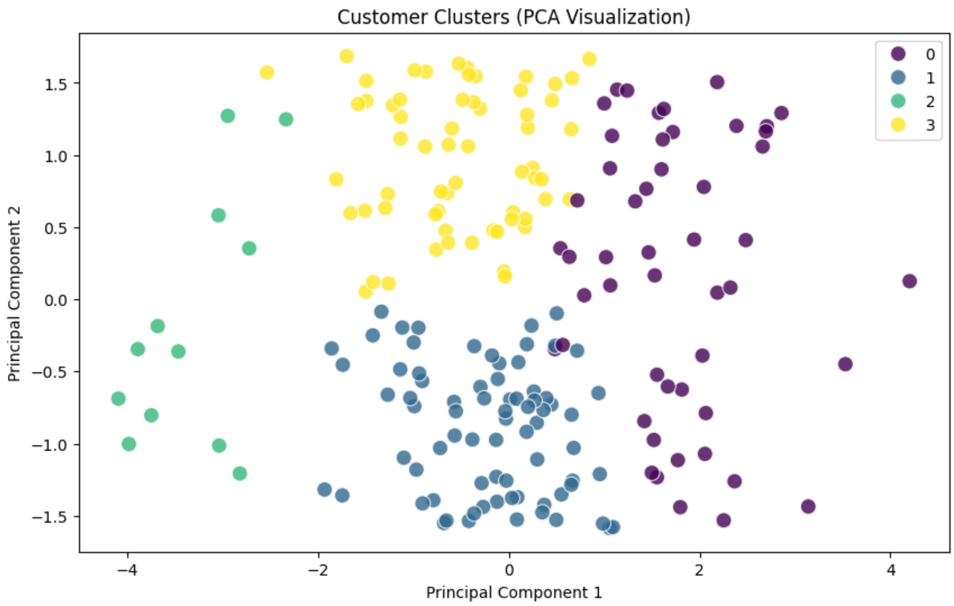

Figure 2: PCA Visualization of Customer Clusters.

3.2 Cluster Profiles

The clusters were analyzed based on key features. Below is a summary:

Cluster	Recency	Frequency	Monetary	Tenure	Region
0	15	10	2000	365	North
1	30	5	800	180	South
2	7	20	3000	500	East
3	45	2	400	90	West

Table 2: Cluster Profiles.

4 Results and Insights

4.1 Clustering Metrics

- **Number of Clusters:** 4
- **Davies-Bouldin Index:** 1.352
- **Silhouette Score:** 0.222

4.2 Insights

- Cluster 0 represents high-value customers with frequent purchases.
- Cluster 3 includes dormant customers with low engagement.
- Targeted marketing strategies can be developed based on cluster characteristics.

5 Code Appendix

Critical snippets from the implementation:

5.1 Feature Engineering

```
def create_clustering_features(df):
    # Code to generate RFM and other features
```

5.2 Clustering Evaluation

```
for k in range(2, 11):
    kmeans = KMeans(n_clusters=k)
    labels = kmeans.fit_predict(data)
    db_scores.append(davies_bouldin_score(data, labels))
```

5.3 Final Clustering

```
optimal_clusters = 4
kmeans = KMeans(n_clusters=optimal_clusters)
cluster_labels = kmeans.fit_predict(data)
```