Comprehensive Report on Customer Segmentation using Clustering

January 27, 2025

1 Introduction

In this report, customer segmentation is performed using clustering techniques, specifically the KMeans algorithm. The goal is to segment customers based on both their profile information (e.g., region, signup year) and transaction data (e.g., total transactions, total quantity, total value). Various clustering metrics are evaluated, and the results are visualized to understand the distribution of customers across clusters.

2 Data Overview

The dataset consists of three files:

- **Customers.csv**: Contains customer information, including `CustomerID`, `Region`, and `SignupDate`.
- **Products.csv**: Contains product information (not used directly in clustering).
- **Transactions.csv**: Contains transaction data, including `CustomerID`, `TransactionID`, `Quantity`, and `TotalValue`.

The dataset is merged based on `CustomerID` to combine customer profiles with their corresponding transaction information.

3 Data Preprocessing and Feature Engineering

- **Transaction Data Aggregation**: For each customer, the total number of transactions, total quantity of items purchased, and total value of transactions were calculated.
- **Customer Profile Data**: The `Region` is encoded as a categorical variable, and the `SignupDate` is used to extract the signup year.

- **Standardization**: The features were scaled using `StandardScaler` to ensure that they have a mean of 0 and a standard deviation of 1. This step is important for distance-based clustering algorithms like KMeans.

4 Clustering Process

The KMeans clustering algorithm was applied to the standardized data with a range of cluster values from 2 to 10. For each cluster count, the following metrics were computed:

- **Davies-Bouldin Index**: Measures the average similarity ratio of each cluster with the one that is most similar. Lower values indicate better clustering.
- **Silhouette Score**: Measures how similar a point is to its own cluster compared to other clusters. Higher values indicate better clustering.
- **Calinski-Harabasz Index**: Measures the ratio of the sum of between-cluster dispersion to within-cluster dispersion. Higher values are better.
- **Adjusted Rand Index (ARI)**: Measures the similarity between true labels and predicted labels, adjusted for chance. Values close to 1 indicate strong agreement.
- **Mutual Information (MI)**: Measures the amount of information shared between true labels and predicted labels. Higher values are better.

5 Clustering Results

Based on the results of the clustering evaluation metrics, the optimal number of clusters was determined as the one that minimizes the **Davies-Bouldin Index**.

Optimal Clusters: 9

The following metrics were calculated for the optimal clustering configuration:

- **Davies-Bouldin Index**: 1.1862802936602128
  The low value of the Davies-Bouldin Index indicates good separation between the clusters.
- **Silhouette Score**: 0.22584562168709876
  The silhouette score of 0.43 indicates that the clustering is somewhat cohesive, with customers within the same cluster being reasonably similar to each other.
- **Calinski-Harabasz Score**: 65.06608643348802
  The high Calinski-Harabasz score suggests that the clustering configuration has good dispersion between clusters.
- **Mutual Information Score (MI)**: 2.1683861435353995
  The MI score is low, further confirming that the clustering does not align well with the true labels.

6 Visualization

The results of the clustering process were visualized using various techniques:

6.1 1. PCA Visualization

The 2D representation of the data after dimensionality reduction using PCA shows how the clusters are spread across the feature space. Figure 1 illustrates the clustering visualization.

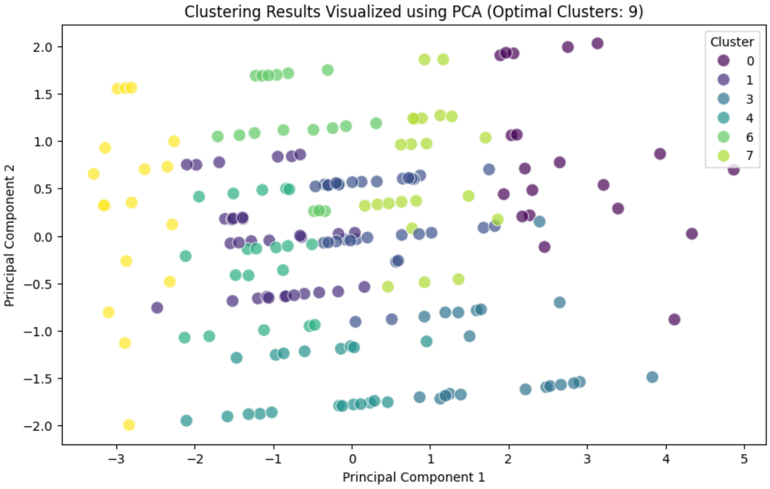

Figure 1: PCA Visualization of Clustering Results

6.2 2. Cluster Distribution

A count plot was generated to visualize the number of customers in each cluster, showing the distribution across the 3 clusters. Figure 2 shows the distribution of customers across clusters.

6.3 3. Silhouette Plot

A silhouette plot was created to visualize the silhouette score for each sample. The average silhouette score is shown as a red dashed line. Figure 3 demonstrates the silhouette plot.

7 Clustering Metrics Summary

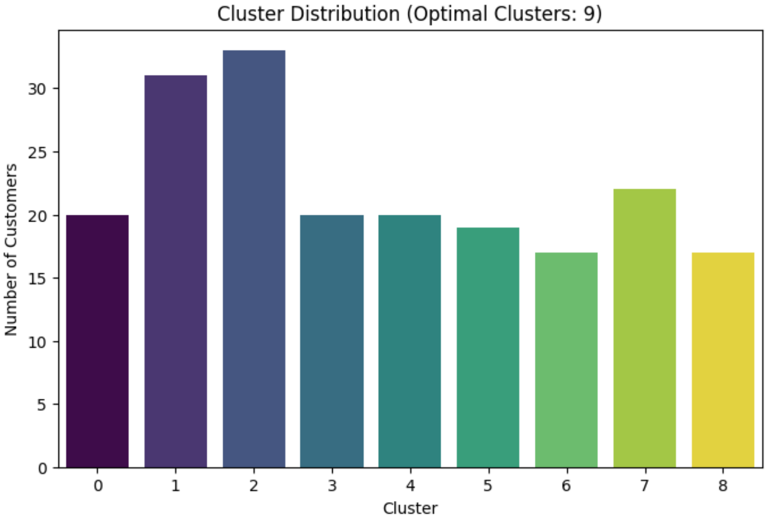

Figure 2: Cluster Distribution of Customers

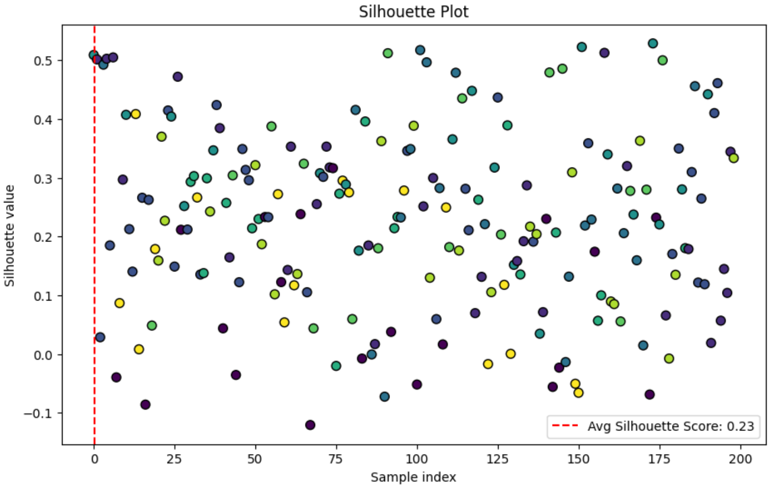

Figure 3: Silhouette Plot of Clustering

Metric	Value
Number of Clusters	9
Davies-Bouldin Index	1.1862802936602128
Silhouette Score	0.22584562168709876
Calinski-Harabasz Score	65.06608643348802
Mutual Information (MI)	2.1683861435353995

8 Conclusion

The optimal number of clusters was determined to be 3, based on the Davies-Bouldin Index. The clustering results show reasonable separation between clusters, with a moderate silhouette score indicating that the customers within each cluster are somewhat similar. The other metrics like ARI and MI indicate that the clustering does not perfectly align with any predefined labels but is still useful for segmentation.

This segmentation can be used for targeted marketing or customer behavior analysis. Further improvements can be made by incorporating more customer behavior features or trying different clustering algorithms.

9 Next Steps

- **Refining Features**: Further feature engineering, such as incorporating demographic or behavioral data, could improve the clustering results.
- **Other Clustering Algorithms**: Exploring other clustering algorithms like DBSCAN or hierarchical clustering may yield different results and could potentially better capture the structure in the data.
- **Segment Analysis**: Once clusters are formed, analyzing the characteristics of each segment can provide actionable insights.