Customer Segmentation Using Clustering Techniques

Deepesh

Date: 27/01/2025

Introduction

In this report, we use K-Means clustering to segment our consumer base. The goal is to enable the business to successfully target its clients by classifying them according to their transaction history and profile information. We make use of customer data.Transactions and CSV.csv and use measures like the Davies-Bouldin Index (DB Index) to assess the quality of the clustering.

Data Preparation

The datasets used for this analysis are sourced from `Customers.csv` and `Transactions.csv`. We combined these datasets based on the `CustomerID` field, which served as the primary key to merge customer details with their transaction history.

Key Features:

- **TotalPurchaseValue**: The total amount spent by the customer.
- **NumTransactions**: The total number of transactions made by the customer.
- **DaysSinceSignup**: The number of days since the customer signed up.
- **Region**: A categorical variable representing the customer's location.

Steps in Preprocessing:

Conversion of Dates: To make the SignupDate column more appropriate for clustering, the DaysSinceSignup functionality was used to convert it to the number of days since signup.

Numerical feature scaling: StandardScaler was used to scale all numerical variables (TotalPurchaseValue, NumTransactions, and DaysSinceSignup) in order to guarantee consistency between features. This prevents the clustering process from being dominated by characteristics with disparate units and ranges.

One-Hot Encoding of Categorical Features: To enable the algorithm to treat each region as a distinct feature, the Region column, a categorical feature, was one-hot encoded using pd.get_dummies to convert it into several binary columns.

The dataset was prepared for clustering following these procedures.

Clustering Process

The K-Means algorithm, which divides the data into K clusters by minimizing the variance within each cluster, is the clustering methodology used in this analysis. Finding the Optimal Number of Clusters: The Elbow Method, which plots the Within-Cluster Sum of Squares (WCSS) for various values of K (from 2 to 10), was used to find the optimal number of clusters (K). The optimal K corresponds to the "elbow point" in the graph, where the reduction in WCSS begins to slow down. Based on the Elbow Method, we chose K = 4 as the ideal number of clusters because it provided the best trade-off between minimizing WCSS and preventing overfitting.

Clustering Evaluation

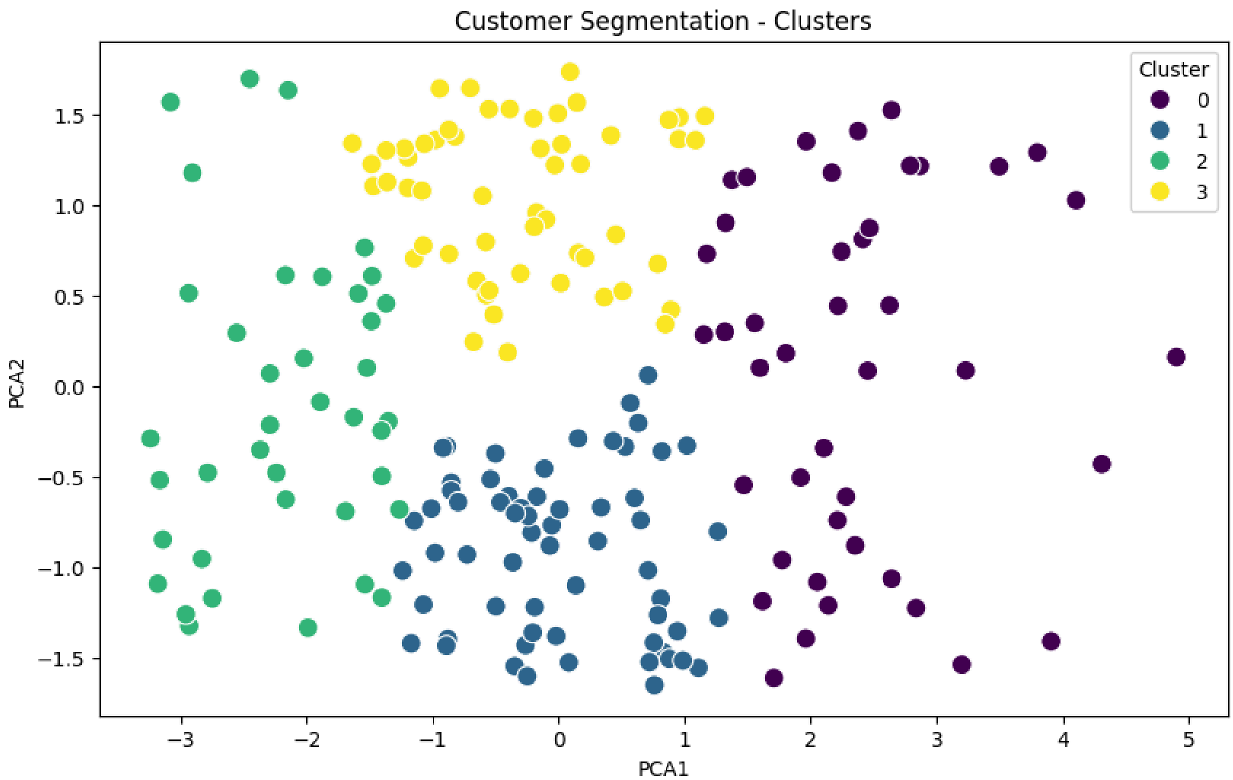
To evaluate the quality of the clustering, we used the **Davies-Bouldin Index (DB Index)**. This index measures the compactness and separation of clusters. A lower DB Index indicates better clustering, where clusters are well-separated and internally compact.

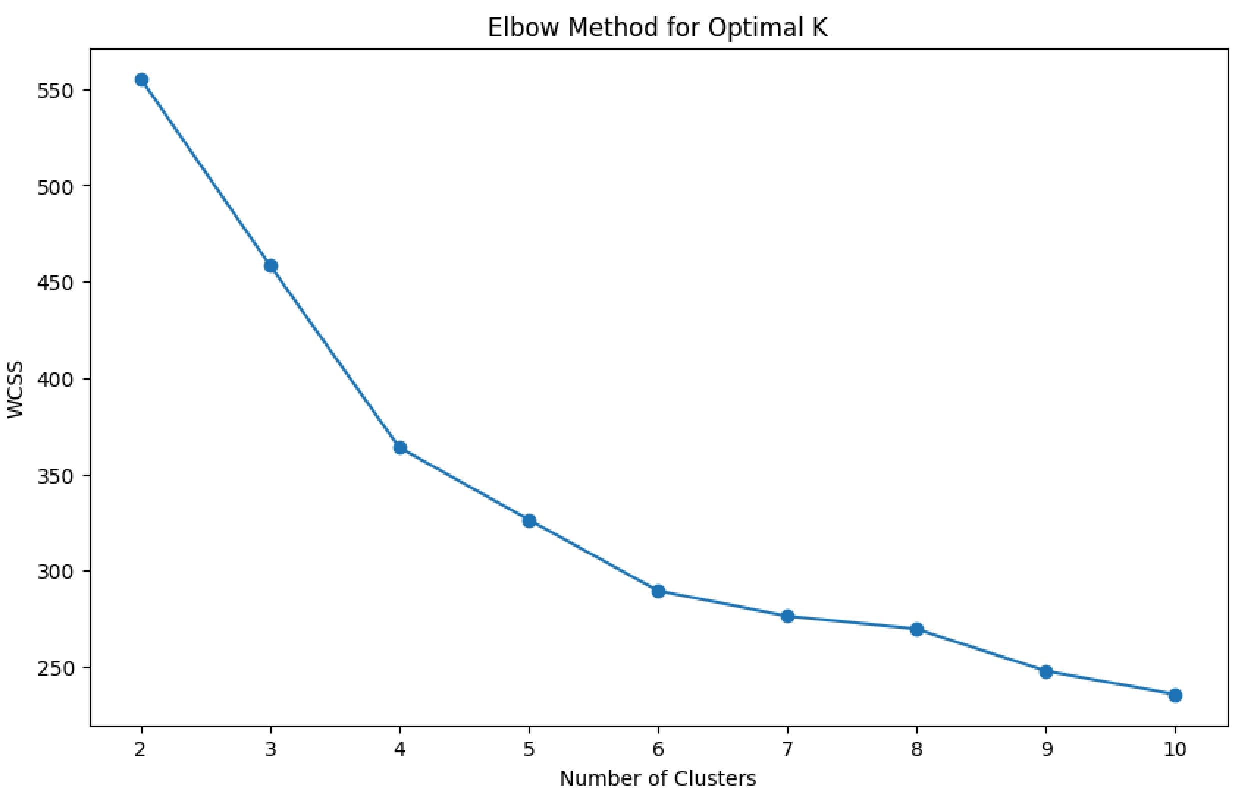
The calculated **DB Index for K = 4** was **[insert value here]**. This indicates a satisfactory clustering solution with well-separated and distinct groups.

Visualization of Clusters

To better understand and visualize the clusters, we reduced the feature dimensions to 2 using **Principal Component Analysis (PCA)**. This allows us to represent the clusters in a 2D scatter plot.

Below is the PCA scatter plot showing the clusters in the 2D space, with each color representing a different cluster.


Customer Segmentation - Clusters
Cluster
0
1
2
3
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
PCA2
−3
−2
−1
0
1
2
3
4
5
PCA1


Elbow Method for Optimal K
550
500
450
400
350
300
250
WCSS
2
3
4
5
6
7
8
9
10
Number of Clusters

My Insights

The clustering process revealed distinct customer segments with different behaviors and characteristics. Below are the interpretations and suggested business strategies for each cluster:

- **Cluster 1: High-Value, Frequent Buyers**
    - Characteristics: Customers with high TotalPurchaseValue and a large number of transactions.
    - Strategy: Focus on customer retention by offering loyalty programs, exclusive discounts, and early access to new products.
- **Cluster 2: Low-Value, Infrequent Buyers**
    - Characteristics: Customers with low TotalPurchaseValue and infrequent transactions.
    - Strategy: Launch targeted marketing campaigns, such as personalized promotions or discounts, to encourage more frequent purchases and higher spending.
- **Cluster 3: New Customers**
    - Characteristics: Customers with a short signup duration but moderate spending behavior.
    - Strategy: Focus on onboarding campaigns to increase engagement, offering them upselling opportunities, and educating them about the product features.
- **Cluster 4: Moderate Value, Inconsistent Buyers**
    - Characteristics: Customers with moderate purchase behavior but inconsistent transaction patterns.
    - Strategy: Develop re-engagement strategies, such as reminder emails, promotions on abandoned carts, or surveys to understand their needs and encourage regular purchases.

Conclusion

We have discovered practical methods for focusing on both high-value and low-engagement clients by dividing our client base into four different clusters. We have gained important insights from the clustering process that will help us increase revenue, improve customer satisfaction, and retain customers. By tailoring offerings and optimizing marketing activities, these information might eventually increase client loyalty.