REPORT

Abhishek Yadav (Student)
[COMPANY NAME] [Company address]

Introduction

Customer segmentation is the practice of dividing a customer base into distinct groups of individuals that share similar characteristics. This enables businesses to target specific segments more effectively and personalize their marketing strategies. In this report, we utilize K-Means Clustering, a popular unsupervised machine learning algorithm, to segment customers based on their purchasing behaviors and demographic details.

Objective

- **Segment the customer base into distinct clusters.**
- **Understand the characteristics of each segment.**
- **Provide actionable insights to the business for targeted marketing.**

Data Description

The dataset used in this report contains customer data with the following attributes:

- **CustomerID**: Unique identifier for each customer.
- **Age**: Age of the customer.
- **Gender**: Gender of the customer.
- **Annual Income** : Annual income of the customer in thousand dollars.
- **Spending Score :** A score assigned by the mall based on customer behavior and spending nature.

Methodology

Data Preprocessing

1. **Handling Missing Values**:
    - Check for any missing values and handle them appropriately.
2. **Encoding Categorical Variables**:
    - Convert the 'Gender' column into numerical format using one-hot encoding.
3. **Feature Scaling**:
    - Normalize the data to ensure that each feature contributes equally to the distance calculations in K-Means.

K-Means Clustering

1. **Optimal Number of Clusters**:
    - Use the Elbow Method and Silhouette Analysis to determine the optimal number of clusters.
2. **Model Training**:
    - Train the K-Means algorithm on the preprocessed data.
3. **Cluster Interpretation**:
    - Analyze the characteristics of each cluster to derive meaningful insights.

Analysis

Data Preprocessing

Handling Missing Values

After examining the dataset, no missing values were found. Therefore, no imputation was necessary.

Encoding Categorical Variables

The 'Gender' column was encoded using one-hot encoding, resulting in two new columns: 'Gender_Male' and 'Gender_Female'.

Feature Scaling

All features were scaled using StandardScaler to ensure they have a mean of 0 and a standard deviation of 1.

Optimal Number of Clusters

Elbow Method

The Elbow Method involves plotting the within-cluster sum of squares (WCSS) against the number of clusters. The point where the WCSS starts to diminish significantly indicates the optimal number of clusters.

Silhouette Analysis

Silhouette analysis measures how similar an object is to its own cluster compared to other clusters. The silhouette score ranges from -1 to 1, with higher values indicating better-defined clusters.

Cluster Interpretation

The characteristics of each cluster were analyzed to derive meaningful insights.

Cluster 1

- **Age Range**: 18-25
- **Income Range**: Low
- **Spending Score**: High
- **Insights**: Young customers with low income but high spending scores. Target for budget-friendly promotions.

Cluster 2

- **Age Range**: 26-35
- **Income Range**: High
- **Spending Score**: High
- **Insights**: Middle-aged customers with high income and high spending scores. Ideal for premium product promotions.

Cluster 3

- **Age Range**: 36-45
- **Income Range**: High
- **Spending Score**: Low
- **Insights**: Older customers with high income but low spending scores. Target with luxury product marketing.

Cluster 4

- **Age Range**: 46-55
- **Income Range**: Low
- **Spending Score**: Low
- **Insights**: Older customers with low income and low spending scores. Target with value-for-money products.

Conclusion

The K-Means clustering algorithm effectively segmented the customer base into five distinct groups, each with unique characteristics. These insights can be leveraged to personalize marketing strategies, enhance customer engagement, and optimize resource allocation. By understanding the specific needs and behaviors of each segment, businesses can deliver more targeted and effective marketing campaigns, ultimately driving higher customer satisfaction and loyalty.