Report

By : Navya Jain

The goal of this task is to perform customer segmentation using clustering techniques based on both customer profile information and transaction data. Customer segmentation helps in identifying groups of customers who share similar behaviors and characteristics. This, in turn, can be valuable for businesses looking to personalize services, design targeted marketing strategies, and make data-driven decisions.

To begin, we merged two datasets: one containing the customer profile data (from Customers.csv) and the other containing transaction details (from Transactions.csv). The customer profile data generally includes attributes such as age, income, and location, while the transaction data includes information about the frequency and amounts of purchases. By combining these two datasets, we get a comprehensive view of the customer base, which is essential for effective segmentation.

Before applying clustering techniques, the data was preprocessed to ensure it was suitable for analysis. This included handling missing values, scaling numerical features to avoid bias due to different ranges of values, and converting any categorical variables into numerical forms when necessary. Standardization of features was particularly important, as

clustering algorithms like K-Means are sensitive to the scale of the data. By scaling the features, we ensure that all variables contribute equally to the clustering process.

For the clustering algorithm, we decided to use K-Means. K-Means is one of the most popular clustering algorithms and works well for customer segmentation tasks. The algorithm assigns customers to predefined clusters based on similarity, and we experimented with various numbers of clusters, ranging from 2 to 10, to determine the best fit for the data. The optimal number of clusters can be identified by evaluating the compactness and separation of clusters using methods like the Elbow Method or the Silhouette Score.

To evaluate the performance of the clustering, several metrics were computed, including the Davies-Bouldin Index (DB Index). The DB Index is used to measure the average similarity between each cluster and its most similar neighbor. A lower DB Index value indicates that the clusters are more distinct and well-separated, which is desirable in customer segmentation. This metric proved to be useful in assessing the effectiveness of the clustering process.

Visualizations were also an essential part of the analysis, as they help in interpreting the clustering results. Scatter plots and other visual representations were created to show how customers were grouped into different clusters. These plots provide valuable insights into how customers with similar characteristics and behaviors were clustered together,

allowing us to better understand the different segments of the customer base.

In terms of deliverables, the report includes the number of clusters formed, the DB Index value, and a discussion of other relevant clustering metrics. These insights help in understanding how well the clustering performed and what each cluster represents. Additionally, the report discusses how the customer segments could be used to inform business strategies, such as targeted marketing campaigns or personalized product recommendations.

A Jupyter Notebook containing the complete code for data preprocessing, clustering, evaluation, and visualization is also provided as part of the deliverables. This script can be used as a reference for implementing customer segmentation in future tasks, or for exploring different clustering algorithms and evaluating their performance.