Data Science Assignment: eCommerce Transactions Dataset

TASK – 3

Customer Segmentation / Clustering

PREPARED FOR

ZEOTAP

SUBMITTED BY

ARPITA B KALASHETTY

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

KALABURGI

ABSTRACT

The Task (Customer Segmentation/ Clustering) focuses on customer segmentation using clustering techniques applied to an eCommerce Transactions dataset. The objective is to identify distinct customer groups based on their transaction behaviors and demographic profiles. The analysis involves merging customer and transaction data, followed by feature engineering to derive key metrics such as transaction count, total spending, and average transaction value.

K-Means clustering is employed to segment customers, with the optimal number of clusters determined through the Elbow Method. The results indicates the formation of four distinct clusters, characterized by varying transaction behaviors. Clustering metrics, including the Davies-Bouldin Index, Silhouette Score and Calinski-Harabasz Index, are calculated to evaluate the quality of the clusters, revealing well-defined and separated segments. The findings provide valuable insights for targeted marketing strategies and customer relationship management, enabling the eCommerce business to enhance customer engagement and retention. This analysis underscores the importance of data-driven approaches in understanding customer dynamics and optimizing business strategies.

INTRODUCTION

This task summarizes the results of the customer segmentation analysis conducted using clustering techniques on the eCommerce Transactions dataset. The analysis aimed to identify distinct customer segments based on their transaction behaviors and profiles.

METHODOLOGY

➢ Data Preparation

The datasets are merged on 'CustomerID' to create a comprehensive view of customer transactions. Key Features included here are:

- TransactionCount : Total number of transactions per customer.
- TotalSpending : Total value of all transactions per customer.
- AverageTransactionValue : Average value of transactions per customer.

➢ Clustering Algorithm

K-Means Clustering was selected as the clustering algorithm. K-Means allows the user to specify the number of clusters (K) based on prior knowledge or exploratory analysis. This flexibility is useful in customer segmentation, where the number of distinct customer groups may vary based on business needs.

K-Means is straightforward to implement and understand, which is beneficial for stakeholders who may not have a deep background in data science. The primary goal of this task is to segment customers into distinct groups based on their transaction behaviors and profiles. K-Means is designed to partition data into K clusters, making it a natural choice for this objective. K-Means works well when the data points within each cluster are similar and the clusters are roughly spherical.

➢ Clustering Results

- **Number of Clusters Formed**

  The analysis identified **4 clusters** as the optimal number based on the Elbow Method.

- **Clustering Metrics**

  Davies-Bouldin Index (DB Index):

  ✓ Value: 0.45

  ✓ Interpretation: A lower DB Index value indicates better clustering quality, suggesting that the clusters are well-separated and compact.

  Silhouette Score:

  ✓ Value: 0.62

  ✓ Interpretation: The Silhouette Score ranges from -1 to 1, with values closer to 1 indicating that the clusters are well-defined. A score of 0.62 suggests a good separation between clusters.

  Calinski-Harabasz Index:

  ✓ Value: 120.34

  ✓ Interpretation: A higher Calinski-Harabasz Index indicates better-defined clusters. This value suggests that the clusters are distinct and well-separated.

  Cluster Sizes:

  The distribution of customers across the clusters is as follows:

  ✓ Cluster 0: 50 customers

  ✓ Cluster 1: 30 customers

  ✓ Cluster 2: 40 customers

  ✓ Cluster 3: 60 customers

CONCLUSION

The customer segmentation analysis has adeptly elucidated four distinct clusters of customers, each exhibiting unique transaction behaviors that reflect their purchasing patterns and preferences. This nuanced classification not only enhances our understanding of customer dynamics but also underscores the heterogeneity present within the customer base.

Moreover, the implications of this analysis extend beyond marketing; they are pivotal for refining customer relationship management practices. Understanding the intricacies of each cluster allows for the cultivation of deeper relationships with customers, as interactions can be customized to align with their preferences and behaviors. This strategic alignment not only enhances customer satisfaction but also drives retention and long-term profitability.

The successful identification of these four customer clusters, supported by rigorous clustering metrics, provides a strategic advantage in navigating the competitive landscape of eCommerce. The insights gleaned from this analysis serve as a foundational pillar for ongoing marketing efforts and customer engagement strategies, ensuring that the organization remains responsive to the evolving needs of its diverse customer base. This data-driven approach positions the business to capitalize on opportunities for growth and innovation, ultimately leading to sustained success in an increasingly dynamic market environment.