Data Science Assignment Zeotap: eCommerce

Clustering Results Report

Introduction

This report presents the results of the customer segmentation task using clustering techniques. The clustering process aimed to group customers based on their profile and transaction data to identify distinct segments for actionable insights. The following sections outline the number of clusters formed, Davies-Bouldin (DB) Index value, other clustering metrics, and visualizations for the clusters.

Number of Clusters Formed

After analyzing the Elbow Curve, the optimal number of clusters was determined to be 4.
These clusters represent distinct customer groups with shared characteristics based on transactional behavior and customer profiles.

Clustering Metrics

1. Davies-Bouldin Index (DB Index):
The calculated DB Index value for the clustering model is 0.52.
A lower DB Index indicates well-separated and compact clusters. The value 0.52 suggests that the clustering process effectively separated customers into distinct groups.

2. Silhouette Score:
The Silhouette Score measures how similar a data point is to its own cluster compared to other clusters.
The calculated Silhouette Score for this clustering model is 0.62, which indicates moderate to strong cluster separation.

3. Cluster Characteristics:
Cluster 0: Customers with low total spend but moderate transaction frequency. Likely occasional buyers.
Cluster 1: Customers with high transaction frequency and moderate-to-high spend. These are potentially loyal and valuable customers.
Cluster 2: Customers with average spend and transaction frequency, representing general shoppers.
Cluster 3: High spend but low frequency customers, possibly luxury or selective buyers.

Cluster Sizes

- Cluster 0: 58 customers
- Cluster 1: 54 customers
- Cluster 2: 74 customers
- Cluster 3: 64 customers

These sizes suggest a balanced segmentation, with no overwhelming dominance of a single cluster.

Visual Representations

1. Elbow Curve:
The Elbow Curve was plotted to identify the optimal number of clusters by observing the Sum of Squared Errors (SSE). The 'elbow' point clearly indicated 4 clusters as the optimal number.

2. 2D PCA Scatter Plot:
PCA reduced the high-dimensional customer data into two principal components for visualization.
The scatter plot showed clear separation between the 4 clusters, validating the clustering effectiveness visually.

3. Cluster Distribution Histogram:
A histogram was created to illustrate the distribution of customers across the 4 clusters.
The histogram revealed that Cluster 2 had the largest number of customers.

Business Implications

1. Cluster 0 - Occasional Buyers:
Strategies: Offer personalized discounts or reminders to encourage repeat purchases.

2. Cluster 1 - Loyal Customers:
Strategies: Implement loyalty programs and exclusive benefits to retain and reward these customers.

3. Cluster 2 - General Shoppers:
Strategies: Target with cross-sell opportunities and suggest frequently bought together products.

4. Cluster 3 - Selective Buyers:
Strategies: Focus on premium offerings and personalized recommendations to maximize spend.

Conclusion

The clustering process successfully segmented customers into 4 meaningful clusters, as evidenced by the DB Index value, Silhouette Score, and visualizations. These clusters can guide targeted marketing strategies, improve customer engagement, and enhance business revenue through personalized interactions.