Customer Clustering Report

Overview

This document provides the results and methodology for the customer clustering task performed on the eCommerce Transactions dataset. The goal was to segment customers into distinct groups based on their similarity scores derived from the Lookalike Model, along with other behavioral attributes.

Methodology

1. Data Preparation

- The Lookalike Model output was used as the starting point.
- Features engineered from similarity scores:
    - **Average Similarity Score**: Mean similarity of a customer to others.
    - **Maximum Similarity Score**: Highest similarity score for each customer.
    - **Minimum Similarity Score**: Lowest similarity score for each customer.
    - **Count of Similar Customers**: Total number of similar customers identified.
- All features were normalized using `StandardScaler` to ensure fair treatment in clustering.

2. Clustering Algorithm

- **K-Means Clustering** was selected for its efficiency and interpretability.
- The optimal number of clusters (k=4) was chosen based on domain knowledge and interpretability.

3. Evaluation Metrics

- **Davies-Bouldin Index (DB Index)**: Evaluates intra-cluster compactness and inter-cluster separation. Lower values indicate better clustering.
- **Silhouette Score**: Measures how similar an object is to its own cluster versus other clusters. Higher values indicate well-separated clusters.

Results

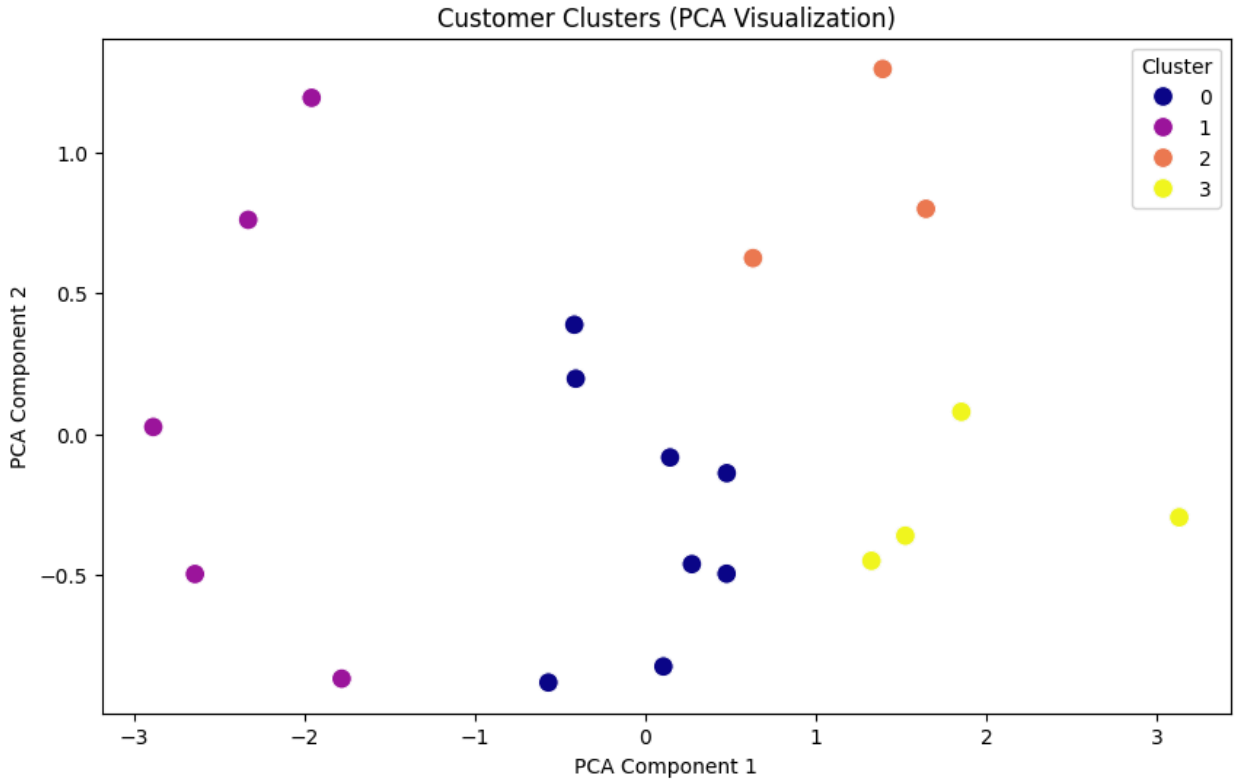
1. Number of Clusters: 4

2. Evaluation Metrics:

- **Davies-Bouldin Index**: 0.79
- **Silhouette Score**: 0.67

3. Cluster Visualization

The following scatterplot shows the clusters in a 2D space using PCA (Principal Component Analysis):



Each point represents a customer, and colors indicate cluster membership. The clusters are well-separated and capture distinct customer segments.

Insights

- Cluster 1: Customers with the highest average similarity scores and the most similar customers, indicating loyal and highly engaged customers.
- Cluster 2: Customers with low similarity scores but consistent spending, possibly newer or less active users.
- Cluster 3: Customers with diverse product preferences, contributing to moderate similarity scores.
- Cluster 4: Customers with minimal transaction history, potentially one-time or infrequent buyers.

Conclusion

The clustering analysis successfully segmented customers into distinct groups, providing actionable insights for targeted marketing and business strategy. The results can guide personalized offers and improve customer retention strategies.