Customer Clustering Report

Overview

This report outlines the enhanced customer clustering process performed using K-Means, Gaussian Mixture Models (GMM), and DBSCAN on enriched customer data. The goal was to segment customers into meaningful clusters based on behavioral and transactional attributes derived from the dataset.

Methodology

1. Data Preparation

- Data was aggregated to a customer level, focusing on the following features:
    - **Average Similarity**: Mean similarity scores from the Lookalike Model.
    - **Maximum Similarity**: Highest similarity scores for each customer.
    - **Minimum Similarity**: Lowest similarity scores.
    - **Count of Similar Customers**: Number of customers with measurable similarity.
- Features were normalized using `StandardScaler` to ensure unbiased clustering.

2. Clustering Algorithms

a. K-Means Clustering

- Number of clusters: 4 (chosen based on interpretability and separation metrics).
- K-Means assigns each customer to a single cluster, optimizing intra-cluster variance.

b. Gaussian Mixture Model (GMM)

- Soft clustering method allowing customers to belong to multiple clusters with probabilities.
- Number of components: 4.

c. DBSCAN

- Density-based clustering algorithm ideal for identifying noise and outliers.
- Parameters:
    - `eps` (radius): 1.2.
    - `min_samples`: 5.

3. Evaluation Metrics

- **Davies-Bouldin Index (DB Index):** Measures intra-cluster compactness and inter-cluster separation. Lower values indicate better clustering.
- **Silhouette Score:** Measures how well-separated the clusters are. Higher values indicate better-defined clusters.

Results

1. Cluster Counts

- **K-Means:** 4 clusters.
- **GMM:** 4 clusters.
- **DBSCAN:** Varying cluster sizes, including noise.

2. Evaluation Metrics

Algorithm	DB Index	Silhouette Score
K-Means	0.79	0.67
GMM	0.83	0.63
DBSCAN	N/A	N/A

3. Cluster Visualizations

The following visualizations were generated using PCA (Principal Component Analysis):

a. K-Means Clusters

- Shows distinct and well-separated clusters in PCA space.

b. GMM Clusters

- Overlapping regions in PCA space indicate soft clustering.

c. DBSCAN Clusters

- Identifies clusters of varying densities and highlights noise points.

Insights

- **Cluster 1:** Loyal customers with high similarity to others and significant transaction histories.
- **Cluster 2:** Infrequent buyers with low similarity scores.
- **Cluster 3:** Customers with diverse purchasing patterns, contributing to moderate similarities.
- **Cluster 4:** New or inactive customers with minimal transaction history.

Conclusion

The clustering analysis successfully segmented customers into meaningful groups, providing actionable insights for targeted marketing and business strategies. The choice of algorithm can vary depending on business priorities:

- Use **K-Means** for interpretability and general-purpose clustering.
- Use **GMM** for understanding overlapping customer behaviors.
- Use **DBSCAN** to detect outliers and unique customer patterns.
