Data Science Assignment: eCommerce Transactions Dataset

Task 3: Customer Segmentation / Clustering

REPORT

Overview

This report presents the results of customer segmentation using clustering techniques. The analysis utilizes both customer profile and transaction information to segment customers into meaningful clusters. The goal is to identify distinct customer groups to enable targeted marketing and enhanced customer engagement strategies.

Data Preparation

Data Loading and Merging

The datasets Customers.csv and Transactions.csv were loaded and merged based on the CustomerID. This step combines customer profiles with their transaction histories, creating a comprehensive dataset for analysis.

Feature Engineering

Features representing customer profiles and their transaction histories were created. These include:

- **Quantity**: Total quantity of products purchased by each customer.
- **TotalValue**: Total monetary value of transactions by each customer.
- **Price**: Average price of products purchased by each customer.
- **ProductID**: Number of unique products purchased by each customer.

These features were normalized to ensure they are on a comparable scale for clustering.

Clustering Algorithm

Elbow Method for Optimal Clusters

The elbow method was used to determine the optimal number of clusters. This method involves plotting the within-cluster sum of squares (WCSS) against the number of clusters and identifying the "elbow point" where the rate of decrease sharply slows down. Based on this method, **4 clusters** were chosen as the optimal number.

K-Means Clustering

The K-Means clustering algorithm was applied with the chosen number of clusters (4).Each customer was assigned to one of these clusters based on their profile and transaction history.

Clustering Metrics

Davies-Bouldin Index (DB Index)

The Davies-Bouldin Index for the chosen number of clusters (4) was calculated. This metric evaluates the quality of clustering by measuring the average similarity ratio of each cluster with respect to its most similar cluster. A lower DB Index indicates better clustering.

- **DB Index**: `1.011942282538582`

Silhouette Score

The Silhouette Score was also calculated to measure how similar each customer is to its own cluster compared to other clusters. This score ranges from -1 to 1, with higher values indicating better-defined clusters.

- **Silhouette Score**: `0.2908762463864909`

Cluster Visualization

The clusters were visualized using scatter plots. Each cluster is represented by a different color, providing a clear visual representation of the customer segments.

```
import seaborn as sns
import matplotlib.pyplot as plt

# Plot the clusters
plt.figure(figsize=(10, 6))
sns.scatterplot(data=customer_features, x='Quantity', y='TotalValue', hue='Cluster',
palette='viridis')
plt.title('Customer Segmentation')
plt.xlabel('Quantity')
plt.ylabel('Total Value')
plt.legend(title='Cluster')
plt.show()
```

Summary

The customer segmentation using clustering techniques provided valuable insights into customer behavior and preferences. By leveraging these clusters, businesses can develop targeted marketing strategies, optimize inventory management, and enhance overall customer experience.