Customer Segmentation Report

Introduction

This report details the customer segmentation analysis conducted to identify distinct customer groups based on their profiles and purchasing behaviors. By leveraging clustering techniques, the objective was to group customers into meaningful segments to facilitate targeted marketing strategies and improved customer relationship management. The analysis utilized two datasets: Customers.csv for customer profile information and Transactions.csv for purchase behaviors. The study also evaluated clustering performance using the Davies-Bouldin Index (DB Index) to identify the most effective algorithm.

Methodology

1. Data Preprocessing

The initial stage involved merging the two datasets and preprocessing the combined data:

- **Feature Aggregation**: Transactional data was aggregated for each customer to calculate:
  - Total spending (TotalValue).
  - Number of transactions (TransactionID count).
  - Total quantity purchased (Quantity).
- **Data Merging**: Aggregated transactional data was merged with the customer profile data, linking by CustomerID.
- **Categorical Encoding**: Categorical variables such as Region were one-hot encoded to allow seamless integration into clustering algorithms.
- **Normalization**: Numerical variables (e.g., TotalValue, TransactionID, Quantity) were standardized using z-scores to ensure all features contributed equally to distance computations.

2. Clustering Algorithms

Four clustering algorithms were evaluated using a grid search for optimal hyperparameters:

- **K-Means**: Focused on finding an optimal number of clusters (n_clusters) between 2 and 10, with variations in initialization methods and maximum iterations.
- **DBSCAN**: Adjusted eps (neighborhood radius) and min_samples to identify dense regions.
- **Agglomerative Clustering**: Experimented with linkage methods (“ward”, “complete”, etc.) and cluster counts.
- **Gaussian Mixture Models**: Tested different numbers of components (n_components) and covariance types.

3. Evaluation Metrics

Clustering performance was assessed using the Davies-Bouldin Index (DB Index). This metric evaluates the compactness and separation of clusters, with lower values indicating better-defined clusters. Additional validation included visual inspection through PCA-based dimensionality reduction.

Results

- **Best Algorithm**: K-Means
- **Best Parameters**:
    - Number of Clusters (n_clusters): 4
    - Initialization Method: “k-means++”
    - Maximum Iterations: 300
    - Random State: 42
- **Evaluation Metrics**:
    - **Davies-Bouldin Index**: 0.78
    - **Number of Clusters Formed**: 4

The K-Means algorithm outperformed other methods due to its ability to create compact and well-separated clusters. DBSCAN’s performance was hindered by the variability in density across the dataset, while Agglomerative Clustering and Gaussian Mixture Models showed moderate success but lacked consistency.

Visualization: Principal Component Analysis (PCA) reduced the feature space to two dimensions for visualization. This revealed four distinct clusters, each representing unique customer groups. These clusters varied in spending patterns, transaction frequency, and purchasing behavior.

Conclusion

The analysis successfully identified four distinct customer segments using the K-Means algorithm. These groups provide actionable insights for targeted marketing campaigns and personalized customer experiences. For instance, high-spending customers can be prioritized for loyalty programs, while customers with low spending but high frequency may benefit from promotional offers.

The study highlights the importance of data preprocessing and parameter tuning in clustering. Future improvements may include incorporating additional features (e.g., demographics, product categories) or exploring advanced methods like deep clustering.
