Zeotap Data Science Assignment

Task 3: PDF Report on Clustering

Customer segmentation is an essential strategy for businesses to group customers based on shared characteristics. In this task, clustering techniques were applied to profile and transaction data to divide customers into meaningful segments. By doing so, businesses can tailor marketing strategies, enhance customer satisfaction, and drive revenue growth.

Number of Clusters Formed: 4

After experimenting with different numbers of clusters (k = 2 to 10), the optimal number of clusters was determined using the elbow method and Davies-Bouldin Index (DB Index). The final model resulted in **4 clusters**, striking a balance between interpretability and segmentation quality.

DB Index Value: 0.8741471963308304

The Davies-Bouldin Index (DB Index), which evaluates the compactness and separation of clusters, was calculated for different cluster configurations. The DB Index for the final clustering solution (4 clusters) was **0.8741471963308304**, indicating well-defined and distinct clusters. Lower DB Index values suggest better clustering, and this result demonstrates that the clustering model effectively captures the diversity in customer behavior.

Other Relevant Clustering Metrics

- **Cluster Centroids**:
    - **Cluster 0**: TotalSpent = 3282.84, AvgSpent = 573.31, TotalTransactions = 5.80, UniqueProducts = 5.77
    - **Cluster 1**: TotalSpent = 1596.86, AvgSpent = 568.72, TotalTransactions = 2.81, UniqueProducts = 2.74

- **Cluster 2**: TotalSpent = 6016.10, AvgSpent = 762.30, TotalTransactions = 8.00, UniqueProducts = 7.61
- **Cluster 3**: TotalSpent = 4116.50, AvgSpent = 1009.74, TotalTransactions = 4.11, UniqueProducts = 4.00

- **Inertia**: The inertia (within-cluster sum of squares) decreased significantly until 4 clusters, after which the reduction was minimal. This aligns with the elbow method and supports the choice of 4 clusters.
- **Cluster Distribution**: The clusters formed were well-distributed with no significant class imbalance, ensuring meaningful segmentation.

Code Explanation

The clustering process involved multiple steps:

1. **Data Preprocessing**
   - Customer data from Customers.csv was merged with transaction data from Transactions.csv to create a unified dataset containing customer profile and behavioral attributes.
   - Missing values were handled using imputation techniques. Numerical features like spending and transaction frequency were scaled using MinMaxScaler to ensure all variables were on the same scale.
2. **Feature Selection**
   - Key features such as total spending, average transaction value, transaction frequency, and time since the last transaction were derived from the dataset to reflect customer behavior and preferences. These features were then standardized.
3. **Clustering Technique**
   - The KMeans clustering algorithm was employed due to its simplicity and efficiency for segmenting customers.
   - The optimal number of clusters was determined using the elbow method (plotting inertia against the number of clusters) and the DB Index.

- After determining 4 clusters, customers were assigned cluster labels.

4. **Cluster Evaluation**
    - Metrics such as DB Index, silhouette score, and inertia were calculated to evaluate the clustering performance.
    - Cluster centers were analyzed to derive meaningful characteristics of each segment.
5. **Visualization**
    - Principal Component Analysis (PCA) was used to reduce the dimensionality of data and visualize clusters in 2D. Scatter plots highlighted distinct groupings, validating the segmentation results.

Insights from Clustering

- **Cluster 0**: Moderate spenders with frequent purchases – Ideal for cross-selling and loyalty programs.
- **Cluster 1**: Low-spending, infrequent customers – Target with introductory offers and incentives to increase engagement.
- **Cluster 2**: High-spending, frequent customers – Prioritize for VIP services and personalized retention strategies.
- **Cluster 3**: High average spend per transaction but fewer purchases – Focus on premium services and high-value product promotions.