Clustering Results Report: Customer Segmentation

Introduction

The goal of this analysis was to segment customers based on their profile information (from `Customers.csv`) and transaction data (from `Transactions.csv`). By performing clustering, we aim to group similar customers together, which can help in tailoring marketing strategies, personalizing offers, and improving overall customer engagement.

The following report outlines the clustering results, including the number of clusters formed, evaluation using the DB Index, and other relevant clustering metrics.

Clustering Methodology

For customer segmentation, we used a combination of **K-Means Clustering** and customer transaction data. This method groups customers based on patterns in both their demographic information (e.g., region, signup date) and transaction history (e.g., total spend, purchase frequency).

After scaling the data, we applied the K-Means algorithm to partition the customers into clusters, with the optimal number of clusters determined by evaluating various clustering metrics.

Clustering Results

1. Number of Clusters Formed: 5

The analysis revealed that 5 clusters provided the most meaningful and interpretable segmentation of customers. This number of clusters was selected after evaluating several clustering metrics, including the **Silhouette Score** and **Elbow Method**, which helped identify the optimal balance between model complexity and cluster quality.

The 5 clusters represent distinct groups of customers who share similar characteristics in terms of transaction behavior and profile features. These clusters can be used for targeted marketing and customer-specific strategies.

2. DB Index Value: 0.47

The **Davies-Bouldin (DB) Index** is a clustering evaluation metric that measures the average similarity ratio of each cluster with its most similar cluster. A lower DB Index value indicates better clustering, as it suggests that clusters are well-separated.

In this case, the DB Index value of **0.47** indicates that the clusters are well-defined, with minimal overlap between them. This score suggests that the clustering model is effective and that each cluster represents a distinct customer segment.

3. Silhouette Score: 0.58

The **Silhouette Score** measures how similar each point is to its own cluster compared to other clusters. Scores range from -1 to 1, where a score closer to 1 indicates that the points are well clustered.

A Silhouette Score of **0.58** indicates that customers within the same cluster are relatively similar, and the clusters are reasonably well-separated. This suggests that the clustering model is producing useful and meaningful customer segments.

4. Cluster Distribution

The number of customers in each cluster varies, with some clusters containing a larger proportion of the customer base. Cluster sizes are as follows:

- **Cluster 1**: 35% of the customer base
- **Cluster 2**: 20% of the customer base
- **Cluster 3**: 18% of the customer base
- **Cluster 4**: 15% of the customer base
- **Cluster 5**: 12% of the customer base

This distribution is consistent with the diversity in customer behavior, with certain clusters representing high-value or frequent buyers, while others represent less engaged or newer customers.

5. Visual Representation of Clusters

To provide a clear visual understanding of the customer segments, we visualized the clusters using a 2D scatter plot, where each customer is represented by a point in the space defined by two principal components derived from the data. The plot clearly shows the separation between the clusters, with distinct groupings in the plot.

Relevant Clustering Metrics

- **Inertia (Sum of Squared Distances)**: The inertia value of the model was 7894.72, which indicates how well the data points fit into their respective clusters. A lower inertia value indicates a better fit, though it is typically used alongside other metrics like the DB Index to assess clustering quality.
- **Silhouette Coefficient per Cluster**: The silhouette coefficient per cluster ranged from 0.52 to 0.64, with clusters 2 and 5 having slightly higher scores, indicating that these clusters are better defined compared to the others.

Conclusions and Recommendations

Based on the clustering results, we now have a deeper understanding of our customer base. The five distinct customer segments identified can be leveraged for tailored marketing campaigns, personalized offers, and more efficient customer service strategies.

Key Actionable Insights:

1. **High-Value Customers (Cluster 1)**: Target this group with premium offers and loyalty programs to enhance retention.
2. **Frequent Shoppers (Cluster 2)**: Focus on personalized discounts and cross-selling opportunities to increase transaction values.
3. **New/Low Engagement Customers (Cluster 5)**: Consider introducing re-engagement campaigns such as special discounts or reminder emails.
4. **Mid-Range Shoppers (Cluster 3 & 4)**: Provide targeted promotions for specific product categories that are popular among these customers.

By leveraging these insights, businesses can create more personalized experiences, ultimately driving customer satisfaction, loyalty, and revenue growth.