TASK 3 REPORT

name: Akshay Goel
email: akshaygoel13062003@gmail.com

Introduction

Customer segmentation is a crucial part of understanding diverse customer behaviors and tailoring marketing efforts accordingly. This report provides an analysis of customer segmentation performed using three clustering algorithms:

KMeans, Agglomerative and Birch.

The dataset used for this analysis was derived from transactional and customer data (excluding product data), as mentioned in the accompanying PDF. Key customer metrics, including account age, total spend, transaction count, frequency, average spend, average quantity, last purchase recency, and region, were derived from the raw transactional and customer data during the preprocessing phase. These derived columns represent vital customer behavior metrics that were crucial in identifying meaningful customer segments.

The analysis leverages PCA (Principal Component Analysis) to reduce the dimensionality of the dataset and uncover key patterns in customer behavior. The clustering algorithms were then applied to segment customers based on their transactional and behavioral characteristics.

The metrics used to evaluate the clustering methods are Davies-Bouldin Index, Silhouette Score, and Calinski-Harabasz Score. These metrics provide a clear picture of the effectiveness of each algorithm in segmenting customers meaningfully.

2. Clustering Methods Used

Three distinct clustering algorithms were applied to the customer data for segmenting the customer base:

KMeans Clustering

KMeans is a centroid-based clustering algorithm that partitions the data into a specified number of clusters, which is predefined by the user. The algorithm works by assigning data points to the closest cluster center (centroid), which is initially chosen randomly. The clusters are refined iteratively by adjusting the centroids based on the mean of the data points assigned to them. This process repeats until the centroids no longer change, or a maximum number of iterations is reached. KMeans minimizes the within-cluster sum of squared distances, making it efficient for finding compact, spherical clusters. It's widely used for its simplicity and efficiency in clustering large datasets.

Agglomerative Clustering

Agglomerative clustering is a hierarchical, bottom-up approach. Unlike KMeans, which requires specifying the number of clusters in advance, Agglomerative Clustering starts by treating each data point as a separate cluster and then iteratively merges the closest clusters based on a distance metric (like Euclidean or Manhattan distance). The merging process continues until the desired number of clusters is achieved or all data points belong to a single cluster. This algorithm does not require the number of clusters to be predefined and produces a tree-like structure called a dendrogram, which visually represents the nested clusters.

Birch Clustering

Birch (Balanced Iterative Reducing and Clustering using Hierarchies) is another hierarchical clustering method specifically designed for large datasets. Unlike traditional hierarchical algorithms, Birch builds a tree structure known as the CF (Clustering Feature) tree, which represents the data points in a compact form. This structure allows Birch to handle large-scale data efficiently. It uses a two-step process: the first step builds the CF tree, and the second step performs clustering using a distance threshold to group the data. Birch is especially beneficial when scalability and performance are crucial for large datasets, as it minimizes the need for excessive memory.

3. Evaluation Metrics

Three evaluation metrics were used to assess the quality of the clusters produced by each algorithm:

- **Davies-Bouldin Index (DBI)**: Measures the average similarity ratio of each cluster with the cluster that is most similar to it. Lower values indicate better clustering.
- **Silhouette Score**: Measures how similar an object is to its own cluster compared to other clusters. Higher values indicate better-defined clusters.
- **Calinski-Harabasz Score**: Measures the ratio of the sum of between-cluster dispersion to within-cluster dispersion. Higher values indicate better separation between clusters.

4. **Clustering Results:**

The analysis suggests that the most suitable number of clusters identified is 4, which provides a balanced segmentation of customers. This choice is based on the evaluation metrics, which indicate that 4 clusters offer meaningful distinctions in customer behavior, preferences, and spending patterns, while also maintaining adequate homogeneity within each cluster.

Metric	KMeans	Agglomerative	Birch
Davies-Bouldin Index	0.744183	0.731018	0.747574
Silhouette Score	0.428076	0.39389	0.384846
Calinski-Harabasz Score	369.364866	362.479822	307.877329

Cluster Profiles:

1. **Cluster 0: Moderate Spend, Low Frequency, Moderate Recency**
   a. **Total Spend**: $225,481.34 (Moderate)
   b. **Frequency**: Low (0.0166)
   c. **Recency**: Moderate (91.67 days)
   d. **Profile**: Customers in this cluster exhibit moderate spending behavior with infrequent purchases and moderate recency. They engage occasionally, making purchases roughly once every 91 days.
   e. **Potential Profile**: A segment of moderately engaged customers with intermittent purchases. They could benefit from re-engagement strategies.
2. **Cluster 1: High Spend, Moderate Frequency, Moderate Recency**
   a. **Total Spend**: $258,909.44 (High)
   b. **Frequency**: Moderate (0.0253)
   c. **Recency**: Moderate (94.08 days)
   d. **Profile**: High-spending customers who make purchases at a moderate frequency. Their recency is slightly higher than Cluster 0, indicating they might be more recently active.
   e. **Potential Profile**: High-value customers with regular spending. They could be targeted for loyalty programs or high-ticket product offerings.
3. **Cluster 2: Low Spend, Very Low Frequency, High Recency**
   a. **Total Spend**: $67,825.14 (Low)
   b. **Frequency**: Very Low (0.0071)
   c. **Recency**: High (136.94 days)
   d. **Profile**: Customers who show very low spending and purchase frequency. They have high recency, suggesting they made a recent purchase but haven't been active for long.
   e. **Potential Profile**: Low engagement customers who could be targeted with reactivation campaigns, potentially offering incentives to bring them back.
4. **Cluster 3: High Spend, Moderate Frequency, Recent Purchases**
   a. **Total Spend**: $137,779.64 (High)
   b. **Frequency**: Moderate (0.0203)
   c. **Recency**: Recent (65.65 days)
   d. **Profile**: These customers spend a lot, with moderate frequency and recent purchases. They show a solid engagement level and are recent buyers, meaning they are highly active.
   e. **Potential Profile**: High-value, active customers, perfect for upselling or introducing premium products.

Summary:

- **Cluster 0**: Infrequent, moderate spenders—ideal for re-engagement strategies.
- **Cluster 1**: High-value customers—engage with loyalty or special offers.
- **Cluster 2**: Low spenders, high recency—target with reactivation campaigns.
- **Cluster 3**: High-value, active customers—focus on upselling and retention strategies.

Insights:

- **High Spenders**: Cluster 1 stands out as the group of high spenders with frequent transactions, indicating that they are highly engaged and bring in significant revenue.
- **Low Spenders & Low Engagement**: Cluster 2 is characterized by low spend and low engagement, which may signal a need for re-engagement strategies to increase their value.
- **Moderate Engagement**: Clusters 0 and 3 have moderate spending and frequency, with Cluster 3 slightly ahead in terms of spending. These customers could benefit from tailored offers or reminders to stay active.

Conclusions

- Effectiveness of Clustering: Use of KMeans, Agglomerative, and Birch clustering algorithms reveals that different customer segments exist that behave differently. The best solution for clustering obtained with 4 clusters gives information on how the customers are dissimilar in their spending, frequency of purchase, and product preferences.

- Customer Segments: Every cluster is a different customer profile. Each one allows for appropriately targeted marketing strategies. For instance, Cluster 0 customers have low spending but high recency. This might indicate that they need some form of re-engagement. A book and home decor-lover - that's what cluster 1 scores high. It makes them an ideal target for promotions of specific products.

- Metrics and Evaluation: The performance of the clustering was evaluated using metrics such as Davies-Bouldin Index, Silhouette Score, and Calinski-Harabasz Score. KMeans performed better than the other algorithms in terms of balancing cluster cohesion and separation, making it the most suitable algorithm for this dataset.

- Business Implications: It is through customer profiling insights that a business can fine-tune marketing, product offerings, and strategies for retaining customers that are suitable for each cluster. For instance, premium offers can be targeted at the high-value clusters, while low-value clusters will require retention and engagement strategies.

- Future Recommendations: It is recommended to continuously monitor and reassess the clusters as new data becomes available. Also, combining clustering with other predictive analytics models could enhance customer insights and drive more effective business decisions.

END