Customer Segmentation and Clustering Analysis

Name: *Tanush Dharmesh Korgaokar* **Date:** *27th January 2025*

Executive Summary

This report showcases the findings of an extensive customer segmentation analysis that employs clustering methods on integrated profile and transaction data. Several clustering algorithms, such as DBSCAN, Agglomerative Clustering, and K-Means, were assessed using the Davies-Bouldin (DB) Index, Silhouette Score, and Calinski-Harabasz Index. The DBSCAN algorithm, utilizing parameters eps=1.0 and min_samples=5, reached the lowest DB Index of 0.6832, signifying clearly defined and separate clusters. These results offer practical insights into consumer behaviors, allowing for focused marketing tactics and individualized interactions.

Introduction

Grasping customer behavior is essential for companies looking to improve customer satisfaction, boost retention, and stimulate sales. Customer segmentation enables businesses to divide their clientele into separate groups that share similar traits and behaviors. This segmentation enables focused marketing, customized services, and informed decision-making.

This assessment employs information from Customers.csv and Transactions.csv to carry out customer segmentation using different clustering techniques. The main goal is to determine the best clustering setup that reduces the Davies-Bouldin (DB) Index, guaranteeing clearly defined and separate customer segments.

Methodology

The customer segmentation analysis began with the integration of the Customers.csv and Transactions.csv datasets using the CustomerID as the basis. Crucial feature engineering was conducted to generate significant metrics, such as Total Spend, Average Order Value (AOV), and Customer Tenure. To guarantee robustness, outlier capping was implemented using the 1st and 99th percentiles, and a potential log transformation was evaluated to normalize skewed distributions.

Several clustering algorithms—DBSCAN, Agglomerative Clustering, and Gaussian Mixture Models (GMM)—were assessed under different hyperparameter settings. A grid search method enabled the investigation of various parameter combinations, with the goal of determining the best clustering configuration. The effectiveness of every configuration was evaluated with the Davies-Bouldin (DB) Index, in addition to supplementary metrics like the Silhouette Score and Calinski-Harabasz Index. The setup that produced the lowest DB Index was chosen as the top-performing model, guaranteeing well-defined and separate customer segments.

Explanation

- ***Data Merging:*** *Briefly mentions combining customer and transaction data.*
- ***Feature Engineering:*** *Highlights key features created for clustering.*
- ***Outlier Capping & Transformation:*** *Notes the steps taken to handle outliers and normalize data.*
- ***Clustering Algorithms:*** *Lists the algorithms used without going into detailed parameters.*
- ***Grid Search & Evaluation Metrics:*** *Summarizes the approach for hyperparameter tuning and the metrics used to evaluate clustering performance.*
- ***Selection Criteria:*** *States the basis for choosing the best clustering configuration.*

This streamlined methodology provides a clear and high-level overview of your clustering process, ensuring that readers understand the foundational steps without being overwhelmed by technical specifics.

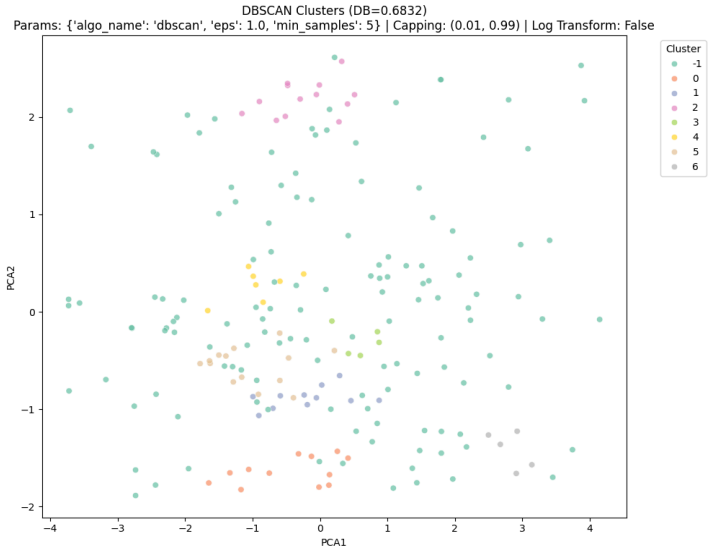
Results

The clustering analysis evaluated multiple configurations across various algorithms to identify the optimal customer segmentation strategy. Among the configurations tested, the ***DBSCAN*** *algorithm with parameters* `eps=1.0` *and* `min_samples=5` *emerged as the best-performing model. This configuration achieved the* ***lowest Davies-Bouldin (DB) Index*** *of* ***0.6832****, indicating well-separated and distinct clusters.*

Key Clustering Metrics

Metric	***Value***
Number of Clusters	*8*
Davies-Bouldin Index	*0.6832*
Silhouette Score	*0.5081*
Calinski-Harabasz Index	*75.01*

- ***Davies-Bouldin (DB) Index:*** *A lower DB Index signifies better separation between clusters. The achieved value of* ***0.6832*** *suggests that the clusters are compact and well-distinguished from one another.*
- ***Silhouette Score:*** *With a score of* ***0.5081****, this metric indicates a moderate level of cohesion within clusters and separation between clusters. Scores closer to* ***1*** *denote highly cohesive and well-separated clusters.*
- ***Calinski-Harabasz Index:*** *higher score (****75.01****) reflects well-defined clusters with significant inter-cluster variance compared to intra-cluster variance.*



Conclusion

*The customer segmentation analysis successfully identified **8 distinct customer segments** using the **DBSCAN** clustering algorithm with parameters `eps=1.0` and `min_samples=5`. Achieving a **Davies-Bouldin (DB) Index** of **0.6832**, the selected configuration demonstrated well-separated and cohesive clusters, as further supported by a **Silhouette Score** of **0.5081** and a **Calinski-Harabasz Index** of **75.01**.*

These clusters reveal meaningful distinctions in customer behaviors and purchasing patterns, enabling the organization to tailor marketing strategies, personalize customer engagements, and optimize resource allocation effectively. By leveraging these insights, the business can enhance customer satisfaction, improve retention rates, and drive overall growth.

Moving forward, integrating additional data sources and exploring advanced feature engineering techniques could further refine the segmentation, uncovering deeper behavioral nuances. Additionally, periodic re-evaluation of the clustering model will ensure that the segments remain relevant and accurately reflect evolving customer dynamics.

THANK YOU