Clustering Evaluation Report

January 27, 2025

1 Introduction

In this analysis, I used K-Means clustering to segment customers based on their profile and transaction history. The objective was to identify distinct customer groups that can inform targeted marketing strategies and enhance customer relationship management. Clustering was performed with the number of clusters (k) ranging from 2 to 10. The quality of each clustering solution was evaluated using four key metrics: Davies-Bouldin Index (DBI), Silhouette Score, Calinski-Harabasz Score, and Inertia.

2 Clustering Evaluation Metrics

The table below summarizes the performance of K-Means clustering for different values of k:

Table 1: K-Means Clustering Evaluation Metrics

k	Davies-Bouldin	Silhouette	Calinski-Harabasz	Inertia
2	2.2538	0.1463	37.4284	2018.4477
3	1.9619	0.1511	32.8729	1799.4575
4	1.6472	0.2410	34.3805	1572.4994
5	1.5246	0.2385	34.5032	1405.3515
6	1.5929	0.2267	32.2481	1310.6622
7	1.6062	0.2054	30.4530	1232.8405
8	1.6527	0.1883	28.6721	1173.3999
9	1.6528	0.1882	27.7841	1109.1941
10	1.6741	0.1825	25.9495	1076.6251

Davies-Bouldin Index (DBI): Lower values indicate better cluster separation.
Silhouette Score: Higher values reflect better-defined clusters.
Calinski-Harabasz Score: Higher scores signify denser and better-separated clusters.
Inertia: Measures how internally tight each cluster is; lower values indicate tighter clusters.

3 Metric Analysis

3.1 Davies-Bouldin Index (DBI)

The Davies-Bouldin Index assesses the average similarity ratio of each cluster with its most similar one. Lower DBI values denote better separation between clusters. Observing the DBI values:

- The DBI decreases from 2.2538 at $k = 2$ to 1.5246 at $k = 5$, indicating improved cluster separation.
- Beyond $k = 5$, the DBI begins to increase, suggesting that additional clusters may lead to over-segmentation.

3.2 Silhouette Score

The Silhouette Score measures how similar an object is to its own cluster compared to other clusters. Higher scores indicate well-defined clusters.

- There is a significant increase from $k = 3$ (0.1511) to $k = 4$ (0.2410), and it remains relatively high at $k = 5$ (0.2385).
- Although $k = 5$ slightly dips compared to $k = 4$, the improvement in DBI suggests that $k = 5$ offers a better balance between cluster separation and cohesion.

3.3 Calinski-Harabasz Score

The Calinski-Harabasz Score evaluates cluster dispersion, with higher values indicating better-defined clusters.

- At $k = 5$, the score reaches 34.5032, slightly higher than at $k = 4$ (34.3805).
- This marginal improvement supports the selection of $k = 5$ as an optimal number of clusters.

3.4 Inertia

Inertia measures the sum of squared distances of samples to their nearest cluster center. Lower values signify tighter clusters.

- Inertia consistently decreases as k increases, which is expected as more clusters typically lead to smaller within-cluster distances.
- However, the rate of decrease slows after $k = 5$, aligning with the DBI and Silhouette findings.

4 Optimal Cluster Selection

Considering all metrics, $k = 5$ emerges as the optimal number of clusters:

- **Lowest DBI**: 1.5246, indicating the best separation among tested values.
- **High Silhouette Score**: 0.2385, suggesting well-defined clusters.
- **Highest Calinski-Harabasz Score**: 34.5032, reinforcing cluster quality.
- **Significant Inertia Reduction**: From 1572.4994 at $k = 4$ to 1405.3515 at $k = 5$.

Final Verdict: Best k (by lowest DB Index) is $k = 5$, with a DB Index of 1.5246.

Clustering the customers into five distinct groups provides a meaningful segmentation that balances clear separation with internal cohesion, facilitating targeted marketing and personalized customer engagement strategies.

5 Conclusion

By selecting $k = 5$ clusters, the company can effectively segment its customer base into distinct groups, each with unique characteristics and behaviors. This segmentation enables the development of targeted marketing strategies, personalized customer experiences, and optimized resource allocation, ultimately driving customer satisfaction and revenue growth.