Customer Segmentation for Starbucks: Clustering Analysis to Increase Sales Among Unloyal Customers

Trixia Nicole A. Libunao
College of Computing and Information Technologies
National University - Manila
Manila, Philippines
libunaota@students.national-u.edu.ph

Jannah Cleine T. Glodo
College of Computing and Information Technologies
National University - Manila
Manila, Philippines
glodojt@students.national-u.edu.ph

***Abstract*— This study provides valuable insights into Starbucks sales analysis by examining consumers' demographic profiles and spending habits. The researchers applied three algorithmic clustering methods — K-Means, Hierarchical Clustering, and Gaussian Mixture Model (GMM) — to segment customers based on their purchasing behavior. These methods were evaluated using key performance metrics, including the Silhouette Score, Davies-Bouldin Score, and Calinski-Harabasz Score, to assess clustering effectiveness and the distinctiveness of customer groups. The findings aim to determine the most suitable algorithm for customer segmentation, offering a deeper understanding of sales patterns and helping Starbucks refine its marketing strategies to enhance customer retention and business growth.**

***Index Terms*—Unsupervised Learning, Clustering Algorithms, K-Means, Hierarchical, Gaussian Mixture Model, Customer Segmentation**

I. Introduction

In recent years, many people have been addicted to the feeling of a nice cozy place to hang out with friends, study, unwind, and even find solitude. A nice café means a well-brewed coffee that invigorates the body and ensures you'll come back for the same strong and bold taste. However, some people visit cafés mainly for the scenery and are not entirely satisfied with the taste of the coffee. Still, they find it acceptable to return, as it is within their budget, and the familiar taste keeps them coming back. Others prioritize how well a barista can extract their go-to coffee, knowing that the quality of the coffee matters more than just having a good ambience—especially if the business's main purpose is not well executed.

A coffee shop's success depends on several factors, including the quality and variety of its products, the atmosphere, location, cultural influences, and customer service. Major chains like Starbucks have built strong brand identities that attract a diverse customer base, but competition remains fierce. Independent coffee shops and other major brands like Dunkin', Tim Hortons, and McCafé continue to challenge Starbucks by offering unique flavors, pricing strategies, and personalized experiences.

Andrew Fender [1], a coffee enthusiast, highlights the difficulty of determining which coffee shop dominates in popularity. Starbucks, founded in 1971 as a small café in Seattle, has since expanded to over 30,000 locations worldwide. Its success is driven by its commitment to quality, an extensive menu, and a strong global presence. The brand's recognizable green mermaid logo is seen in cities and small towns alike, reinforcing its dominance in the industry. However, despite this widespread presence, customer loyalty varies, and Starbucks still faces challenges in retaining certain customer segments. This study aims to analyze customer segmentation at Starbucks using clustering techniques to better understand purchasing patterns, particularly among unloyal customers. By identifying distinct customer profiles, the study seeks to uncover key factors contributing to customer retention and disloyalty. These insights will help Starbucks develop targeted marketing strategies, improve customer satisfaction, and ultimately increase sales within these less-engaged customer groups [1].

II. Review of Related Literature

Customer segmentation is a fundamental concept in marketing and business analytics, aiming to categorize customers based on shared characteristics. Traditional segmentation approaches relied on demographic, psychographic, and behavioral data, but modern techniques increasingly leverage machine learning models to enhance precision.

Clustering techniques, such as K-Means, Hierarchical Clustering, and DBSCAN, are widely employed in customer segmentation. These unsupervised learning methods group customers based on similarities in purchasing behavior, demographics, or engagement levels. Principal Component Analysis (PCA) is often used alongside clustering to reduce dimensionality and improve feature interpretability [2].

Historically, customer segmentation was based on simple rule-based classifications, such as RFM (Recency, Frequency, Monetary) analysis. With advancements in computing power and data availability, businesses have transitioned to AI-driven approaches that incorporate behavioral data, transactional history, and external factors such as seasonality [2].

Starbucks has leveraged advanced machine learning techniques to optimize customer targeting and promotional strategies. A study [2] applied K-Means clustering and PCA to Starbucks' customer dataset, identifying distinct customer groups based on spending behavior and promotional response. The findings revealed that feature engineering, such as incorporating demographics, transaction history, and offer types, significantly improved clustering effectiveness.

Another study on Starbucks segmentation [3] demonstrated how businesses can use data-driven methods to allocate promotions more effectively. By analyzing customer spending habits and response rates, they tailored marketing campaigns to enhance engagement and revenue growth.

III. Methodology

A. *Data Collection*

The dataset was retrieved from Muhammet Gamal's "Starbucks Satisfactory Survey Analysis" on Kaggle. This dataset includes demographics, purchase patterns, factors influencing purchases, and reasons for spending. These attributes provide critical insights into customer behavior, allowing for meaningful segmentation. However, the dataset lacks clarity regarding the origin of spending values, which are denominated in Malaysian currency.

B. *Data Pre-Processing*

The dataset required multiple preprocessing steps to ensure accuracy and compatibility with clustering algorithms. The following steps were taken:

1. Handling Missing Values – Missing values in numerical columns were replaced with the mean of their respective features.
2. Feature Encoding – Categorical variables were transformed into numerical representations. This included converting labels like *Income, Spending Behavior, and Loyalty* into numerical values.
3. Feature Selection – Features most relevant to Starbucks customer behavior were retained, including factors such as price sensitivity, promotion influence, and service ratings.
4. Feature Scaling – Standardization was applied using StandardScaler, ensuring all features had a mean of 0 and a standard deviation of 1. This prevented features with larger ranges from dominating the clustering results.

C. *Experimental Setup*

The experiment was conducted in Google Colab using a Python notebook. The tools and libraries were utilized:

- Scikit-learn: For modeling and preprocessing.
- Pandas: For data handling and manipulation.
- Matplotlib: A Python library for creating visualizations.
- NumPy: For numerical processing.
- Standard Scaler: For scaling data to unit variance.
- Principal Component Analysis (PCA): To simplify data by identifying key features.
- K-means, Hierarchical Clustering, Gaussian Mixture Model (GMM): Clustering techniques.

D. *Algorithm*

Three algorithmic models were applied to the dataset:

1. K-Means Algorithm – Chosen for its efficiency and ability to segment customers into meaningful groups. The Elbow Method was used to determine the optimal number of clusters (x=5).
2. Hierarchical Clustering (Agglomerative) – This algorithm was used for comparison, providing insights into the hierarchical relationships between customer segments.
3. Gaussian Mixture Model (GMM) – This probabilistic model was included as an alternative to K-Means to evaluate whether soft clustering would provide a better understanding of overlapping customer behaviors.

E. *Training Procedure*

The training procedure involved clustering and analysis to evaluate results. Key steps include:

1. Applying PCA – Dimensionality reduction was performed to allow visualization in 3D space and remove redundant information.
2. Training the Clustering Models – The dataset was clustered using K-Means, Hierarchical Clustering, and GMM.
3. Cluster Profiling – After clustering, the mean values of features in each cluster were analyzed to understand segment characteristics.
4. Customer Segmentation Interpretation – The clusters were examined to identify patterns in spending behavior, loyalty, and promotional influence.

F. *Evaluation Metrics*

Evaluated the models using the following metrics:

1) Silhouette Score – Measures how well-separated clusters are. A higher score indicates better-defined clusters.
2) Davies-Bouldin Score – Evaluates the compactness and separation of clusters. Lower values indicate better clustering.
3) Calinski-Harabasz Score – Measures the variance between and within clusters. Higher values indicate more distinct clusters.

G. Comparison of Clustering Algorithms

- K-Means was selected as the primary clustering method due to its efficiency and interpretability. Hierarchical Clustering was used for benchmarking, and GMM was added to see if a soft clustering approach would yield better results.
- The Silhouette Score ranked K-Means as the best-performing algorithm, followed by Hierarchical Clustering, with GMM performing the lowest.
- While GMM was expected to provide better cluster flexibility, its lower Silhouette Score suggests that customer segments are more distinct than overlapping.

IV. Results and Discussion

1) Customer Segmentation Outcomes

The application of K-Means clustering successfully segmented Starbucks customers into distinct groups based on spending patterns, preferences, and behavioral factors. The analysis identified five primary clusters:

1. High-Spending Loyal Customers – Regular customers who frequently purchase premium Starbucks products and have high brand loyalty.
2. Budget-Conscious Consumers – Customers who prioritize affordability and tend to buy lower-cost menu items.
3. Occasional Shoppers – Individuals who visit Starbucks infrequently, often influenced by promotions.
4. Ambiance-Oriented Visitors – Customers who visit primarily for the café environment rather than the coffee itself.
5. Unloyal or Irregular Consumers – Those who rarely visit Starbucks and are least likely to return.

2) Model Evaluation Metrics

The three clustering algorithms (K-Means, Hierarchical Clustering, and Gaussian Mixture Model) were evaluated using Silhouette Score, Davies-Bouldin Score, and Calinski-Harabasz Score to determine how well the models segmented customer data.

Clustering Method	Silhouette Score	Davis-Bouldin Score	Calinski-Harabasz Score
K-Means	0.290521	1.094197	55.306251
Hierarchical	0.221555	1.180685	46.596756
GMM	0.133505	1.538040	22.005415

Table 1. Clustering Evaluation Metrics

- Silhouette Score: Measures how well-defined the clusters are (higher is better). K-Means performed best.
- Davies-Bouldin Score: Lower values indicate better separation between clusters. Hierarchical performed best.
- Calinski-Harabasz Score: Measures variance within clusters (higher is better). K-Means was superior.

3) Comparative Analysis of Clustering Methods

K-Means perform best among the three, because:

- it is effective in segmenting well-separated groups, which aligns with the characteristics of Starbucks' customers. The results suggest that customer segments are distinct rather than overlapping, making K-Means preferable over GMM.
- Hierarchical Clustering performed slightly worse but provided a strong alternative, as seen in its lower Davies-Bouldin Score. This suggests that some hierarchical relationships exist within customer behavior patterns.
- GMM performed the worst in all three-evaluation metrics. This suggests that soft clustering was not effective for this dataset because Starbucks customers tend to belong distinctly to one segment rather than being probabilistically distributed across multiple clusters.

4) Interpretation of Key Results

- The clustering models successfully grouped Starbucks customers based on spending habits and loyalty.
- The findings support the hypothesis that Starbucks can target specific customer segments with personalized marketing strategies.
- The results validate previous assumptions that customer behavior is highly dependent on pricing, promotions, and brand loyalty.

Emerging Patterns and Trends

- Loyalty and Spending Correlation – Higher-spending customers showed strong loyalty, while low-spending customers were less committed to the brand.
- Influence of Promotions – Customers in lower loyalty segments were often influenced by seasonal discounts or limited-time promotions.
- Ambiance-Driven Customers – A subset of visitors cared more about store atmosphere and experience than coffee pricing or taste.

5) Alignment with Previous Research

- Previous research on customer segmentation in the food & beverage industry has shown that brand loyalty and spending behavior are closely linked.
- Similar clustering methods have been used in retail and hospitality industries, supporting the effectiveness of K-Means for consumer segmentation.
- However, unlike other studies where soft clustering (GMM) performed well, this study found clear, distinct customer groups, making hard clustering more effective.

6) Limitations and Future Improvements

This research highlights valuable insights for Starbucks' marketing strategies, emphasizing the importance of customer segmentation in enhancing business performance.

1. Limited Dataset Scope – The dataset only covers a specific period and location, limiting generalizability to Starbucks globally.
2. Lack of Temporal Data – The study does not account for seasonal changes in customer behavior.
3. Potential for Hybrid Clustering Approaches – Combining Hierarchical Clustering with K-Means could provide even more refined customer segments.

For our future work:

- Incorporating time-series data to analyze seasonal trends in Starbucks customer behavior.
- Exploring deeper feature selection methods to improve cluster separability.

Conducting a comparative analysis with additional clustering algorithms such as DBSCAN (if improved), Mean-Shift, or Spectral Clustering.

7) The study successfully applied K-Means clustering to segment Starbucks customers based on spending patterns, brand loyalty, and purchasing behavior. The findings indicate that Starbucks can improve customer retention and increase sales by targeting specific customer groups with personalized promotions. K-Means was confirmed as the best clustering model**,** outperforming both Hierarchical Clustering and GMM.

V. Conclusion

This study effectively segmented Starbucks customers into distinct groups based on their spending patterns, brand loyalty, and motivations for visiting the café. The results demonstrate that while Starbucks' global success is built on its strong brand identity and quality offerings, customer loyalty remains a complex dynamic influenced by pricing strategies, promotions, and the overall café experience. Utilizing three algorithmic models, the K-Means algorithm proved to be the most effective clustering method, revealing clear and well-separated customer segments — ranging from high-spending loyal patrons to ambiance-seekers and irregular consumers.

By understanding these segments, Starbucks can tailor its marketing strategies, enhance customer satisfaction, and address gaps in customer retention. For instance, targeted promotions may attract budget-conscious consumers, while refining store ambiance could further appeal to atmosphere-driven visitors. Additionally, incorporating temporal data and exploring hybrid clustering techniques could provide even deeper insights into evolving consumer behaviors.

Ultimately, the study highlights the importance of data-driven decision-making in a highly competitive industry. By leveraging customer segmentation, Starbucks can strengthen its market position, foster long-term loyalty, and continue adapting to the diverse needs of its customer base worldwide.

VI. References

[1] Fender, A, Is Starbucks the most popular coffee shop? - KitchenSurfers. KitchenSurfers, April 2018, https://kitchensurfers.com/is-starbucks-the-most-popular-coffee-shop/

[2] Seifip, M. (2021). Starbucks offers advanced customer segmentation with python. Retrieved from https://seifip.medium.com/starbucks-offers-advanced-customer-segmentation-with-python-737f22e245a4

[3] Malakar, A. (2021). Starbucks customer segmentation promotion with k-means and pca. Retrieved from https://medium.com/@malakar.abh/starbucks-customer-segmentation-promotion-with-k-means-and-pca-adb9be8f4e43

[4] EdrawMind. (2023). Starbucks Market Segmentation, Targeting, and Positioning. Retrieved from https://www.edrawmind.com/article/starbucks-market-segmentation-targeting-and-positioning.