**Starbucks Customer Segmentation Using K-Means for Effective Marketing Strategies**

**Mike Rasell Carale Dago-oc1, Chris Angelu Bongcawil Jordan2**

Computer Science and Information Technology Department,

Bachelor of Science in Computer Science (BSCS)

Negros Oriental State University (NORSU), Negros Oriental, Philippines

|  |  |  |
| --- | --- | --- |
| **Article Info** |  | **ABSTRACT** |
| ***Keywords:***  *Clustering*  *K-Means*  *K-Means Elbow Method*  *Principal Component Analysis*  *Silhouette Score*  *Davies-Bouldin Score* |  | This study applies K-Means clustering to segment customers based on demographic, income, and membership behavior data. Using a dataset of 1,700 instances with four selected features: age, membership duration, income, and gender, the research identified that three clusters represent the optimal segmentation. This conclusion was supported by the Elbow Method, which revealed diminishing Within-Cluster Sum of Squares (WCSS) beyond cluster 3 and validated by clustering quality metrics, where cluster 3 achieved the highest Silhouette Score (0.431) and the lowest Davies-Bouldin Score (0.809). The clustering analysis revealed distinct customer segments projected onto two principal components through PCA: High-Income Customers, Middle-Income Customers, and Lower Income Customers. These clusters provide actionable insights into customer behavior, enabling businesses to design targeted marketing strategies, enhance customer engagement, and optimize resource allocation. The findings demonstrate the effectiveness of combining dimensionality reduction and clustering techniques in customer segmentation for informed decision-making. |
|  |

1. **INTRODUCTION**

Clustering is a critical machine learning technique used to group data points based on inherent similarities, enabling businesses to uncover meaningful insights from complex datasets. In the context of consumer behavior, clustering offers the potential to identify distinct customer segments, each characterized by unique spending patterns, preferences, and purchasing habits. This segmentation allows organizations to tailor their offerings and marketing strategies to meet the needs of specific customer groups, enhancing overall customer satisfaction and business outcomes. This study focuses on applying clustering techniques to analyze customer income data and its influence on purchasing behaviors. By identifying clusters such as "Low Spender Customer", "Medium Spender Customer", and "High Spender Customer", businesses can design targeted promotions, optimize pricing, and develop loyalty programs that cater to the unique characteristics of each segment. These insights have practical applications in marketing, product development, pricing strategies, and customer retention, enabling data-driven decisions to improve profitability and brand loyalty [1].

1. **LITERATURE REVIEW**

**2.1. Clustering in Customer Behavior Analysis**

Clustering has been extensively employed in customer behavior analysis, providing insights into spending patterns, preferences, and purchasing habits. In particular, clustering helps businesses segment customers into distinct groups based on income, location, or other demographic factors. For example, income-based clustering has been used to study spending behavior, enabling businesses to identify high-value customers and their preferences [2]. Such segmentation aids in designing targeted marketing campaigns and enhancing customer engagement strategies [3].

**2.2. Popular Clustering Algorithms in Consumer Analysis**

Widely-used clustering methods like K-Means, DBSCAN, and Hierarchical Clustering have proven effective in analyzing customer data. K-Means is commonly employed for partitioning customers into income-based clusters, as it minimizes intra-cluster variance and ensures distinct segmentation [4][5]. DBSCAN excels in identifying outliers, such as occasional big spenders or customers with irregular purchasing habits [6]. Hierarchical Clustering is ideal for uncovering nested relationships, such as those between income groups and product preferences [7].

**2.3 Applications of Clustering in Income-Based Segmentation**

Income-based segmentation has been widely applied in retail and service industries to enhance marketing effectiveness and business growth. For instance, K-Means clustering has been utilized to group customers by income, revealing patterns in spending habits and product preferences. This approach enables businesses to design targeted promotions, such as loyalty rewards for regular customers or premium offerings for high-income segments. Additionally, DBSCAN has been employed to detect outliers in spending behavior, uncovering niche markets and unique customer needs [8].

**2.4 Evaluation Metrics for Clustering Algorithms**

Metrics like the Elbow Method, Silhouette Score, and Davies-Bouldin Index are essential for evaluating clustering performance. The Elbow Method determines the optimal number of income-based clusters by analyzing reductions in intra-cluster variance [9]. The Silhouette Score assesses the cohesion and separation of clusters, providing a measure of how well each customer fits within their assigned group [10]. The Davies-Bouldin Index evaluates the compactness and distinctness of clusters, with lower values indicating better-defined groupings [11].

**2.5 Insights from Income-Based Clustering in Marketing**

Clustering customer income data provides actionable insights for targeted marketing and pricing strategies. For instance, luxury-oriented customers often prioritize premium products and exclusive benefits, while budget-conscious segments respond to affordability and discounts [12]. By understanding these preferences, businesses can optimize their offerings to maximize customer satisfaction and revenue. Clustering also informs location-based decisions, guiding businesses on store placement or regional marketing campaigns tailored to specific income groups [13].

1. **METHODOLOGY**

**3.1. Materials**

**3.1.1. Datasets**

This research utilizes a dataset containing 1,700 instances and six features, of which four were selected for analysis. These features are age, became\_member\_on, income, and gender, providing demographic and temporal insights into customer characteristics. The two unused features, feature 1 (ID number) and ID were excluded as they did not contribute to the analysis. The selected dataset enables the evaluation of patterns in Starbucks customer demographics and membership behavior, forming the foundation for clustering and identifying trends in user profiles [14].

**3.1.2. Hardware**

The study was conducted on a system running Windows 10 with a 64-bit operating system. The hardware specifications included an Intel Core™ i5-7400 CPU and 16GB of Random Access Memory (RAM).

**3.1.3. Software**

The researcher utilized Jupyter and Python version 3.9.15 as the primary programming environment for data analysis and model implementation. The following libraries were employed: OS, NumPy, Pandas, Seaborn, Matplotlib, and Scikit-Learn.

**3.2. Methods**

**3.2.1. Data Gathering**

This data was gathered by Starbucks to simulate their customers and transactions to see if there are better approaches to sending customers specific promotional deals.

**3.2.2. Data Pre-processing**

Data processing is an essential aspect of model development. Data acquired in their raw form contain noise and anomalies, which can affect the performance and training process of the model being schooled [15]. The researchers employed several techniques to clean the data, which included data normalization and fixing missing values:

Data normalization is a pre-processing technique primarily intended to manage numerical features and is applied to numerical features before the application of classification algorithms. Normalization is crucial to prevent the effect of certain features from being concealed by others, particularly when the ranges of the features are inconsistent [16].

A missing value is a datum that has not been stored or gathered due to issues like faulty sampling procedures, budgetary constraints, or limitations in the data collection process. Missing values are an inevitable aspect of data analysis and can present significant challenges for data practitioners. It is generated due to several reasons, including human mistakes, technical malfunctions, unavailable data, or outdated and inconsistent data [17].

**3.2.3. Principal Component Analysis (PCA)**

Since the dataset has 18 features PCA reduces the number of features or dimensions in the data into 2 features while retaining the most important patterns or variance in the dataset. The remaining features would be principal components 1 and 2, PCA achieves this by transforming the original features into new, uncorrelated variables called principal components. These components are linear combinations of the original features, ordered such that the first principal component captures the maximum variance in the data, followed by the second, and so on [18][19].

**3.2.4 K-Means Clustering**

K-means clustering was employed to partition the dataset into distinct groups based on consumer characteristics. The algorithm aims to minimize intra-cluster variance by iteratively adjusting cluster centroids and assigning data points to their nearest cluster. The Elbow Method was used to determine the optimal number of clusters by analyzing the Within-Cluster Sum of Squares (WCSS) values for different cluster counts. Mathematically, the objective function for K-Means is defined as [20]:

(1)

Where is the total number of clusters, the group of data points in the i-th cluster, a single data point, and the center of the i-th cluster.

**3.2.5 Silhouette Score**

To evaluate the quality of the clusters generated by the K-Means algorithm, the Silhouette Score was calculated. This metric measures how well each data point fits within its assigned cluster compared to other clusters. The score ranges from -1 to 1, where a value closer to 1 indicates that clusters are well-separated and cohesive. A score near 0 suggests overlapping clusters, and negative values indicate that points are assigned to the wrong clusters [21].

The Silhouette Score measures how well a data point fits into its cluster. It is calculated by , the average distance from the point to all other points in the same cluster, and b, the average distance from the point to all points in the nearest cluster.

(2)

**3.2.6 Davies-Bouldin Score**

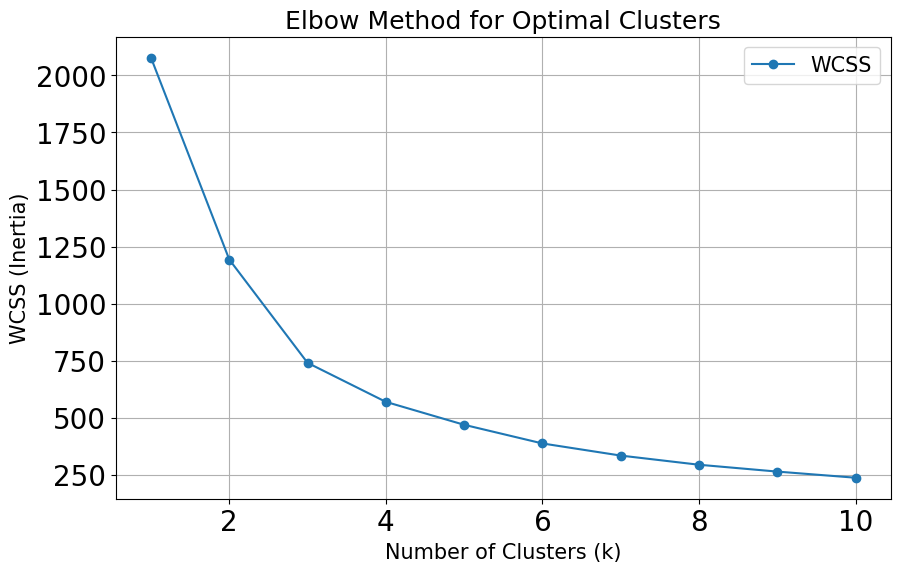
To evaluate the quality of the clusters generated by the K-Means algorithm, the Davies-Bouldin Score was calculated. This metric measures the average similarity ratio between each cluster and its most similar neighboring cluster. Lower Davies-Bouldin Scores indicate better-defined clusters, where each cluster is compact and distinct from others [22].

(3)

Where represents the number of clusters, the average distance of all points in cluster to the centroid of cluster , and is the distance between the centroids of clusters and .

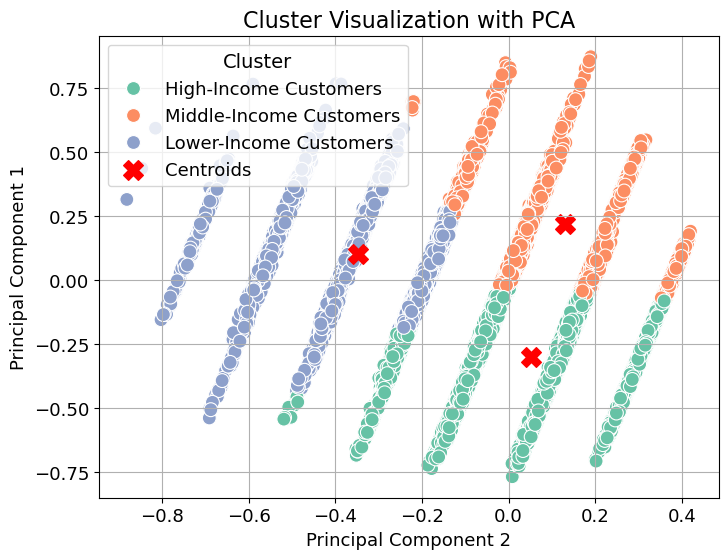
1. **RESULTS AND DISCUSSION**

**4.1. Elbow Method for Optimal Clusters**

****

**Figure 1.** Elbow Method

The curve initially shows a steep decline in WCSS as the number of clusters increases, particularly between 1 and 3 clusters. After this point, the reduction becomes more gradual, forming a visible "elbow" at k = 3. This point is where the decrease in WCSS becomes less steep, indicating diminishing returns for adding more clusters, and indicates that three clusters are the optimal solution.

****

**Figure 2.** Starbucks Customer Segmentation

Figure 2 is a scatter plot that illustrates the data projected onto two principal components, with each point representing an individual customer and colors corresponding to three distinct clusters. Cluster 1, represented in green and located toward the bottom-right, comprises High-Income Customers, indicating individuals with a consistent spending pattern and limited reliance on cash advances. Cluster 2, marked in orange and situated in the upper right region, represents Middle-Income Customers. This group includes regular spenders who make modest purchases and rely moderately on cash advances. Cluster 3, shown in blue and positioned in the center-left, identifies Lower-Income Customers, who balance regular spending with occasional cash advances. The clusters are distinct and well-separated, with minimal overlap between them. This indicates that the clustering algorithm did a good job of grouping similar data points based on the underlying features. The centroids are positioned away from each other, which reflects a strong distinction between the groups. These clusters provide actionable insights into customer behavior and spending patterns, enabling businesses to develop personalized marketing strategies, improve financial planning, and enhance customer engagement. This clustering approach underscores the value of data-driven methods in targeted business applications.

**Table 1.** Cluster Validation Metrics

|  |  |  |  |
| --- | --- | --- | --- |
| Index | Number of Clusters | Silhouette Scores | Davies-Bouldin Scores |
| 0 | 2 | 0.394 | 1.020 |
| 1 | 3 | 0.431 | 0.809 |
| 2 | 4 | 0.383 | 0.870 |
| 3 | 5 | 0.390 | 0.862 |
| 4 | 6 | 0.386 | 0.860 |
| 5 | 7 | 0.394 | 0.851 |
| 6 | 8 | 0.369 | 0.889 |
| 7 | 9 | 0.375 | 0.857 |
| 8 | 10 | 0.386 | 0.810 |

A higher Silhouette Score indicates better-defined clusters, while a lower Davies-Bouldin Score suggests better clustering quality. Table 1 presents the Silhouette Scores and Davies-Bouldin Scores for different numbers of clusters, ranging from 2 to 10. The Silhouette Score is highest at k = 3 (0.431), indicating that the data is most naturally separable into three groups. As the number of clusters increases beyond 3, the Silhouette Score generally decreases, suggesting reduced cluster cohesion and separation with more clusters. After approximately 6 clusters, the score stabilizes, but remains lower than at k = 3, indicating diminishing returns in cluster quality as more clusters are added. The Davies-Bouldin Score is lowest at k = 3 (0.809), further confirming that this configuration produces the best-defined clusters. Although the Davies-Bouldin Score fluctuates slightly beyond 3 clusters, it remains relatively low, suggesting minor improvements in clustering quality with additional clusters. Overall, the k = 3 configuration is the best choice, providing a balance between interpretability and clustering quality based on the combined evaluation of Silhouette and Davies-Bouldin Scores.

**CONCLUSION**

This study successfully applied the K-Means clustering algorithm to segment Starbucks customers into distinct groups based on behavioral and transactional data. The analysis determined that the optimal number of clusters was k = 3, as identified using the Elbow Method and further validated with the highest Silhouette Score (0.431) and the lowest Davies-Bouldin Score (0.809). The clustering results revealed key patterns, such as a smaller but distinct segment of High-Income Customers (Cluster 1), characterized by consistent spending patterns and minimal reliance on cash advances, presenting opportunities for targeted high-value offerings and premium services. Middle-income customers (Cluster 2), who are regular spenders with modest purchases and moderate reliance on cash advances, highlight the importance of budget-friendly marketing strategies. Lastly, Lower-Income Customers (Cluster 3), who balance regular spending with occasional cash advances, offer opportunities for flexible financial products and personalized engagement. The clusters are distinct and well-separated, with minimal overlap, indicating effective grouping of similar data points. Additionally, the centroids' clear separation reflects strong group distinctions. This clustering approach validates the efficacy of K-Means as a reliable tool for uncovering meaningful customer segments. By leveraging these insights, businesses can move beyond one-size-fits-all strategies to adopt data-driven, personalized marketing and engagement initiatives. The findings highlight actionable opportunities, such as budget-friendly offerings for cost-conscious customers and premium services for high-income earners. Overall, this study demonstrates how K-Means clustering transforms raw customer data into strategic insights, enabling businesses to enhance marketing efficiency, improve customer engagement, and optimize resource allocation effectively.

**REFERENCES:**

[1] Liu, B. (2012). Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers.

[2] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1-135.

[3] Archak, N., Ghose, A., & Ipeirotis, P. G. (2011). Deriving the pricing power of product features by mining consumer reviews. Management Science, 57(8), 1485-1509.

[4] Hawkins, R. P., Kreuter, M., Resnicow, K., Fishbein, M., & Dijkstra, A. (2008). Understanding tailoring in communicating about health. Health Education Research, 23(3), 454-466.

[5] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational Linguistics, 37(2), 267-305.

[6] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of KDD (pp. 226-231).

[7] Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. ACM Computing Surveys, 31(3), 264-323.

[8] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (pp. 281-297).

[9] Thorndike, R. L. (1953). Who belongs in the family? Psychometrika, 18(4), 267-276.

[10] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20(1), 53-65.

[11] Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1(2), 224-227.

[12] Kotler, P., & Keller, K. L. (2016). Marketing Management. Pearson.

[13] Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. IEEE Intelligent Systems, 28(2), 15-21.

[14] I. Muliar, "Starbucks Customer Data," Kaggle, Jan. 2019. [Online]. Available: https://www.kaggle.com/datasets/ihormuliar/starbucks-customer-data/data?select=profile.csv. [Accessed: Jan. 27, 2025].

[15] C. I. Eke, A. A. Norman, L. Shuib, and H. F. Nweke, “A Survey of User Profiling: State-of-the-Art, Challenges, and Solutions,” IEEE Access, vol. 7, pp. 144907–144924, 2019, doi: 10.1109/ACCESS.2019.2944243.

[16] M. Gomaa, "Starbucks Customer Data Analysis," Data, vol. 6, no. 2, p. 11, Feb. 2021. [Online]. Available: https://www.mdpi.com/2306-5729/6/2/11. [Accessed: Jan. 27, 2025].

[17] G. Theophilus and C. I. Eke, "Machine Learning-based E-Learners’ Engagement Level Prediction using Benchmark Datasets," International Journal of Applied Information Systems (IJAIS), vol. 12, no. 41, pp. 23, Sept. 2023. [Online]. Available: www.ijais.org. [Accessed: Jan. 27, 2025].

[18] "Lesson 11: Principal Components Analysis (PCA)," Penn State Eberly College of Science. [Online]. Available: https://online.stat.psu.edu/stat505/lesson/11/11.1. [Accessed: Jan. 27, 2025].

[19] "Principal Component Analysis (PCA)," GraphPad Prism. [Online]. Available: https://www.graphpad.com/guides/prism/latest/statistics/stat\_pca\_process\_pc\_defined.htm. [Accessed: Jan. 27, 2025].

[20] "An Effective and Efficient Algorithm for K-Means Clustering with New Formulation," IEEE Journals & Magazine, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/9723527. [Accessed: Jan. 27, 2025].

[21] Y. Guo, Y. Liu and Y. Huang, "Customer segmentation for personalized recommendation in e-commerce using machine learning", Journal of Physics: Conference Series, vol. 1462, no. 1, pp. 012015, 2020.

[22] Lahcen Abidar, Dounia Zaidouni and Abdeslam Ennouaary, "Cus-tomer Segmentation With Machine Learning: New Strategy For Tar-geted Actions", Proceedings of the 13th International Conference on Intelligent Systems: Theories and Applications, 2020.