A Machine Learning Approach to Credit Card Customer Segmentation for Economic Stability

Yujuan Qiu[1,a], Jianxiong Wang[2,b]

[*a] Corresponding author: juliaqiuyj@gmail.com, [b]jeremyjianxiong.wang@gmail.com

[1]School of Engineering and Applied Science, The George Washington University, Washington, D.C., U.S.
[2]Investment Banking, J.P. Morgan, New York, U.S.

Abstract: Credit card usage is a vital component of the global economy, but unpredictable customer behavior poses significant challenges. Machine learning (ML) has emerged as a powerful tool for customer segmentation in the credit card industry. This paper systematically examines different clustering algorithms to identify the most effective approach for accurately categorizing credit card customers. The evaluation of clustering model accuracy is conducted through the Davies-Bouldin Index, Silhouette Score, and Calinski-Harabasz Index. This systematic approach aims to advance our understanding of how ML can be optimally harnessed to enhance customer segmentation, ultimately contributing to economic stability and growth.

Keywords: machine learning; customer segmentation; credit card usage; economic stability

1. Introduction

Financial development is closely linked to a country's resilience, productivity, and growth, and the utilization of credit cards within a nation's economy is not merely a financial transaction but a catalyst for economic growth and stability [1-3]. Credit card usage is closely entwined with consumer spending, a key driver of demand for goods and services across various sectors. However, the unpredictability of customer behavior in credit card usage poses significant challenges. Unforeseen fluctuations in spending can trigger inflation, interest rate changes, and economic instability. Additionally, traditional credit assessment methods may exclude individuals with limited credit histories, hindering financial inclusion and access to credit.

The diversity of customer behavior in credit card usage is both a source of opportunity and a complex puzzle. Understanding the intricacies of customer segments is paramount for financial institutions seeking to optimize their offerings, risk management, and overall strategies [4]. However, this task is far from straightforward. Identifying customer segmentation is a critical challenge in credit card usage, as it directly impacts the ability of financial institutions to tailor products and strategies to diverse customer profiles. Accurately identifying and understanding these segments is essential for promoting economic stability and growth. However, the complexity and unpredictability of customer behavior in credit card usage pose significant obstacles.

ICEMBDA 2023, October 27-29, Tianjin, People's Republic of China

DOI 10.4108/eai.27-10-2023.2342007

Machine learning (ML) has emerged as a powerful tool in many fields [5], especially refining customer segmentation in the realm of the credit card industry, yielding a multitude of advantages [3]. Numerous studies have demonstrated the potential of ML in enhancing customer segmentation within the credit card industry. For instance, Rachman et al. exemplified customer segmentation through the RFM (Recency, Frequency, Monetary) analysis using machine learning clustering, amalgamating it with segmentation based on demographic, geographic, and behavioral data through data warehouse-based business intelligence [6]. Dawood et al. explored the application of various clustering techniques. This innovation reduced clustering execution time and delivered superior accuracy results, ultimately establishing neural networks as the most effective clustering technique [7]. However, it is essential to underscore that the efficacy of machine learning models is intrinsically linked to the quality of the data on which they are trained. Financial institutions must diligently ensure that their machine learning models are trained on high-quality data that faithfully represents their customer base.

This paper refocuses its attention on addressing the intricacies of precise customer segmentation within the credit card industry, driven by the overarching goal of empowering financial institutions to make informed, data-driven decisions that bolster economic stability. To accomplish this objective, a range of clustering algorithms, including K-means, hierarchical clustering, DBSCAN, Birch, and Gaussian mixture, are systematically examined to identify the most effective approach for accurately categorizing credit card customers. The evaluation of clustering model accuracy is conducted through the employment of metrics such as the Davies-Bouldin Index, Silhouette Score, and Calinski-Harabasz Index. This systematic approach aims to advance our understanding of how machine learning can be optimally harnessed to enhance customer segmentation, ultimately contributing to economic stability and growth.

2. Methodology

2.1 Dataset

This research utilizes a dataset that focuses on customer segmentation for the development of targeted marketing strategies within the context of credit card usage. The dataset in question comprises records of approximately 9,000 active credit card holders, capturing their behavioral patterns over the last six months. Each record is at the customer level and includes 18 distinct behavioral variables, enabling a comprehensive analysis of credit card usage patterns and customer segmentation. The 18 distinct behavioral variables are explained in the following:

(1) CUST_ID: A categorical variable representing the unique identification of each credit card holder.

(2) BALANCE: This variable signifies the balance amount remaining in the account available for making purchases.

(3) BALANCE_FREQUENCY: A continuous variable measuring the frequency with which the account balance is updated. It is scored between 0 and 1, with 1 indicating frequent updates and 0 indicating infrequent updates

(4) PURCHASES: This variable represents the total amount of purchases made from the credit card account.

(5) ONEOFF_PURCHASES: Indicates the maximum purchase amount made in a single transaction.

(6) INSTALLMENTS_PURCHASES: The amount of purchases made in installments.

(7) CASH_ADVANCE: Represents cash advances taken by the cardholder.

(8) PURCHASES_FREQUENCY: This variable quantifies the frequency of purchase transactions, with a score between 0 and 1, where 1 indicates frequent purchases and 0 indicates infrequent purchases.

(9) ONEOFFPURCHASESFREQUENCY: Measures the frequency of purchases made in one-go (i.e., single transactions), with a score of 1 indicating frequent occurrences and 0 indicating infrequent occurrences.

(10) PURCHASESINSTALLMENTSFREQUENCY: Reflects how often purchases are made in installments, with 1 denoting frequent installment purchases and 0 indicating infrequent installment purchases.

(11) CASHADVANCEFREQUENCY: Measures the frequency of cash advances being taken by the cardholder.

(12) CASHADVANCETRX: The number of transactions involving "Cash in Advance."

(13) PURCHASES_TRX: Represents the number of purchase transactions made using the credit card.

(14) CREDIT_LIMIT: Denotes the credit limit assigned to the user for the credit card.

(15) PAYMENTS: Signifies the total amount of payments made by the user.

(16) MINIMUM_PAYMENTS: Represents the minimum amount of payments made by the user.

(17) PRCFULLPAYMENT: Indicates the percentage of the full credit card balance paid by the user.

(18) TENURE: Reflects the tenure of the credit card service for the user.

This dataset provides a comprehensive view of customer behavior in the context of credit card usage, encompassing variables related to balance, purchase behavior, credit limit, payment patterns, and more. It serves as a valuable resource for understanding customer segmentation and developing data-driven marketing strategies to enhance economic stability and growth within the financial sector.

2.2 Machine Learning Approach

In our pursuit of accurate customer segmentation for credit card usage, we turn to machine learning techniques, which offer a data-driven and adaptive approach to categorizing customers. Several machine learning algorithms can perform customer segmentation effectively, including K-means [8], hierarchical clustering [9], DBSCAN [10], Birch [8], and Gaussian mixture [8].

For this study, we have chosen to employ the K-means clustering algorithm as our primary method for customer segmentation. We have compared K-means algorithm with other clustering models. The K-means algorithm aligns with our objective of obtaining interpretable and

actionable customer segments to inform strategy decisions. K-means clustering is a partitioning algorithm that aims to divide a dataset into K distinct, non-overlapping clusters, with each cluster represented by a central point called a centroid. The key idea behind K-means is to minimize the sum of squared distances between data points and their assigned centroids, effectively grouping similar data points together.

In leveraging K-means clustering, our aim is to achieve precise and meaningful customer segmentation that informs data-driven strategies within the financial sector. By accurately identifying customer segments, we intend to contribute to the economic stability and growth of the industry, ultimately fostering a more resilient and prosperous financial landscape.

2.3 Evaluation Metrics

In this paper, we employ three key metrics, namely the Davies-Bouldin Index, Silhouette Score, and Calinski-Harabasz Index, to rigorously assess the accuracy and effectiveness of the clustering models.

The Davies-Bouldin Index is a commonly employed metric in machine learning to assess clustering quality. It gauges the average similarity between each cluster and its most akin cluster while taking into account cluster size. The Silhouette Score, on the other hand, measures the degree of resemblance between an object and its designated cluster (cohesion) in comparison to other clusters (separation). The Silhouette Score spans from -1 to +1, where a higher value signifies a strong alignment of the object with its own cluster and a notable dissimilarity from neighboring clusters. The Calinski-Harabasz Index, also referred to as the Variance Ratio Criterion, calculates the ratio between the aggregate dispersion among clusters and the dispersion within clusters for all clusters. A higher Calinski-Harabasz Index score indicates superior performance in clustering.

3. Results and discussion

3.1 Customer Behavioral Overview

Our initial analysis focused on exploring the distribution of critical customer behavioral variables, including balance, purchases, cash advances, and payment patterns. This analysis served as a foundational step in our research, enabling us to identify key parameters that exhibit distinctive characteristics.

Within the dataset, we observed a noteworthy feature—a marked skewness in the distributions of these key variables, particularly in variables such as "PURCHASES" and "CASH_ADVANCE," which displayed significant rightward tails. This skewness highlighted the dynamic nature of credit card usage, with the potential for abrupt changes in spending habits, cash advance utilization, and payment behaviors.

With the identification of key parameters that exhibit distinctive characteristics, our research now advances to a critical phase—analyzing the correlations between these parameters. This phase aims to uncover meaningful relationships and dependencies that can further inform our customer segmentation strategy.

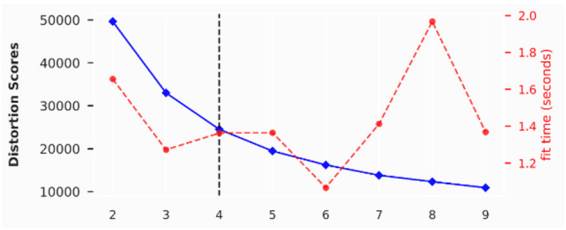
In this analysis, understanding these correlations is essential for creating nuanced customer segments that not only capture shared behavioral characteristics but also account for potential interactions between key parameters. After analysis, it can be seen that the correlation coefficients of the variables before two by two are less than 0.7, indicating that multicollinearity did not occur and the 18 features in the dataset were appropriate [11].

3.2 Initialization of the Center Point

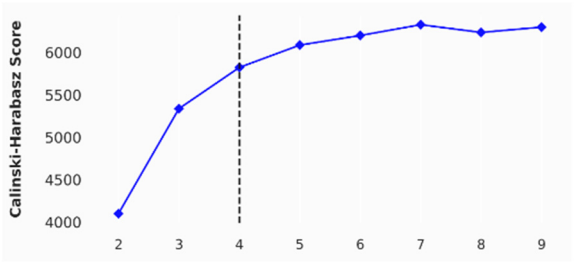
Now we would want to implement machine learning techniques on the sample data based on our observation so far. In real-world applications, applying the K-means algorithm introduces complexities due to its sensitivity to initial random centroid initialization. Unlike controlled scenarios where data clusters are well-defined, real-world data often lacks such clear boundaries. Consequently, K-means may occasionally initialize centroids in a manner that results in suboptimal clustering. To address this challenge and determine the optimal number of clusters ('n'), we employ two widely-used techniques:

- Elbow Method.
- Calinski-Harabasz Index.

The elbow method involves plotting the number of clusters ('n') against the corresponding WCSS values. As 'n' increases, WCSS tends to decrease, as each data point can be closer to its centroid. However, the elbow point in the plot indicates the point where the rate of decrease in WCSS significantly slows down. This 'elbow' suggests an optimal number of clusters where further partitioning provides diminishing returns. Additionally, we leveraged the Calinski-Harabasz Index, also known as the Variance Ratio Criterion. This index measures the ratio of between-cluster variance to within-cluster variance. A higher Calinski-Harabasz score implies better separation between clusters. We calculated this index for various cluster counts, as shown in Fig. 1. As shown in Fig. 1, Upon plotting the number of clusters against Within Cluster Summation of Squares (WCSS) values, we observed a distinct "elbow" in Fig. 1. This point suggested that the rate of WCSS reduction significantly slowed down at 'n' around 3 or 4 clusters. The elbows calculated using both the Elbow Method and the Calinski-Harabasz Index method were for 4. Based on the results obtained from both the Elbow Method and the Calinski-Harabasz Index, it has been determined that the most suitable number of clusters for the K-Means algorithm is 4 clusters.



(a)



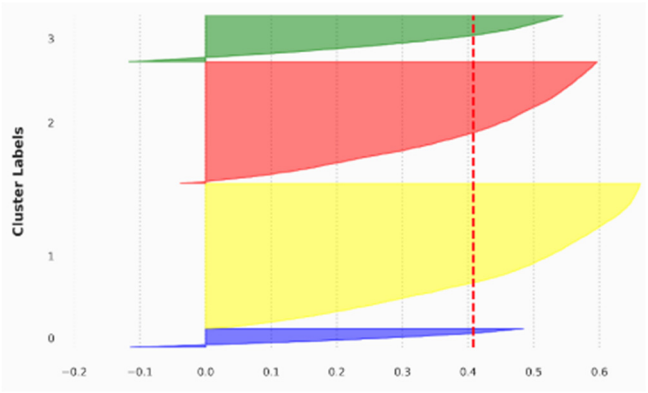
(b)

Fig. 1. K-means clustering optimal number estimation; (a) the Elbow Method and (b) the Calinski-Harabasz Index method.

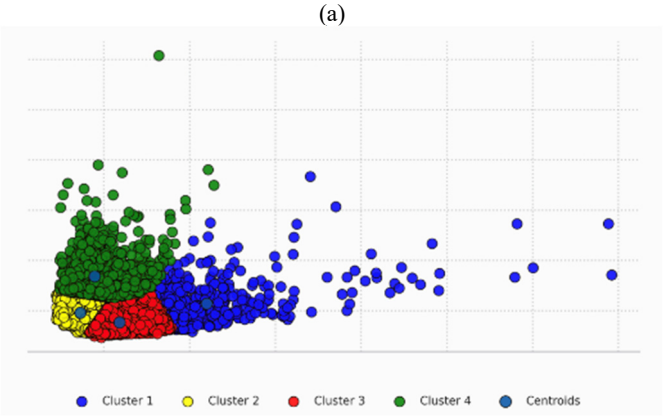
3.3 Results and Discussion of Clustering

Fig. 2 provides a comprehensive representation of cluster analysis, incorporating silhouette coefficient values, cluster distributions within a scatter plot, and individual customer memberships within each cluster. Silhouette coefficients indicate that all clusters exhibit values surpassing the average, affirming their optimal separation. Furthermore, uniform fluctuations in silhouette plot size are observed across all clusters, indicating consistent data point dispersion. However, distinctions arise in terms of thickness, with clusters 2 and 3 displaying notably thicker consistency than others. This heightened thickness in clusters 2 and 3 is attributed to their high viscosity, stemming from a majority of data points concentrated in the bottom-left corner of the scatter plot, where both clusters account for over 35% of the total customer distribution. The K-Means algorithm assigns data outliers to clusters 1 and 4, with x-axis outliers assigned to cluster 1 and y-axis outliers to cluster 4. A pie chart at the visualization's bottom succinctly portrays the percentage distribution of customers within each cluster, as shown in Fig. 2(c).

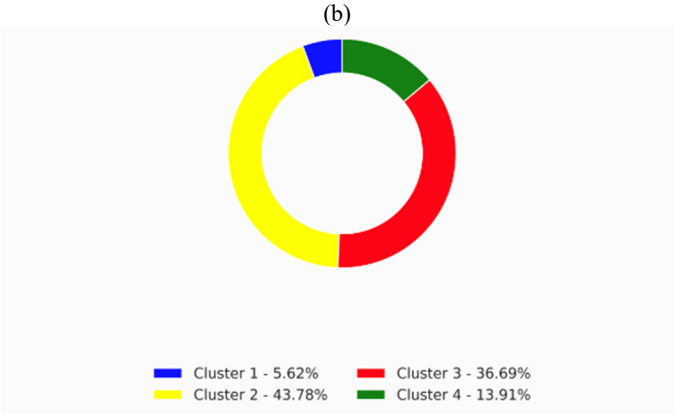
In addition to the K-means clustering method, four representative unsupervised learning algorithms, which are the hierarchical clustering [9], DBSCAN [10], Birch [8], and Gaussian mixture [8], were compared with the K-means clustering method. Furthermore, we use Davies-Bouldin Index, Silhouette Score, and Calinski-Harabasz Index to evaluate the accuracy of the clustering models. The calculation results are shown in Table 1. Our analysis reveals that the K-Means clustering model exhibits commendable clustering quality, underscoring its effectiveness in segmenting the customer dataset. The application of cluster profiling serves as a valuable tool for identifying the distinctive characteristics inherent to each cluster.



(a)



(b)



(c)

Fig. 2. K-means clustering optimal number result: (a) Silhouette plots of clusters, (b) scatter plot clusters distributions, and (c) percentage of each cluster.

Table 1. Accuracy of different clustering models

Method	Davies-Bouldin Index	Silhouette Score	Calinski-Harabasz Index
K-means clustering	0.80	0.87	5823
DBSCAN	0.90	0.43	3749

Birch	0.85	0.64	4531
Gaussian mixture	0.87	0.76	4760
Hierarchical clustering	0.92	0.39	3307

4. Conclusions

This research underscores the invaluable role of machine learning techniques in enhancing customer segmentation within the credit card industry, ultimately contributing to economic stability. Our findings demonstrate the commendable clustering quality of the K-means model, validating its efficacy in segmenting the customer dataset. The dynamic approach to customer segmentation enriches the financial sector's decision-making processes, adapting to ever-evolving customer behaviors. In conclusion, this study underscores the transformative potential of machine learning in credit card segmentation, offering data-driven insights that foster economic stability.

References

[1] A Kumar, AK Mishra, VK Sonkar, S Saroj. "Access to credit and economic well-being of rural households: Evidence from Eastern India." Journal of Agricultural and Resource Economics 45, no. 1 (2020): 145-160.

[2] Y Qiu. Financial Deepening and Economic Growth in Select Emerging Markets with Currency Board Systems: Theory and Evidence. No. 87. The Johns Hopkins Institute for Applied Economics, Global Health, and the Study of Business Enterprise, 2017.

[3] Y Qiu. "Estimation of tail risk measures in finance: Approaches to extreme value mixture modeling." Johns Hopkins University, 2019.

[4] YC Tsao, PVRP Raj, V Yu. "Product substitution in different weights and brands considering customer segmentation and panic buying behavior." Industrial Marketing Management 77 (2019): 209-220.

[5] Z Zhang, H Rui. "Byzantine-Robust Federated Learning with Variance Reduction and Differential Privacy." arXiv preprint arXiv:2309.03437 (2023).

[6] FP Rachman, H Santoso, A Djajadi. "Machine learning mini batch K-means and business intelligence utilization for credit card customer segmentation." International Journal of Advanced Computer Science and Applications 12, no. 10 (2021).

[7] EAE Dawood, E Elfakhrany, FA Maghraby. "Improve profiling bank customer's behavior using machine learning." IEEE Access 7 (2019): 109320-109327.

[8] Y Liu, Y Bao. "Automatic interpretation of strain distributions measured from distributed fiber optic sensors for crack monitoring." Measurement 211 (2023): 112629.

[9] J Guan, S Li, X He, J Zhu, J Chen. "Fast hierarchical clustering of local density peaks via an association degree transfer method." Neurocomputing 455 (2021): 401-418.

[10] Y Chen, L Zhou, S Pei, Z Yu, Y Chen, X Liu, J Du, N Xiong. "KNN-BLOCK DBSCAN: Fast clustering for large-scale data." IEEE transactions on systems, man, and cybernetics: systems 51, no. 6 (2019): 3939-3953.

[11] Y Liu, Y Bao. "Real-time remote measurement of distance using ultra-wideband (UWB) sensors." Automation in Construction 150 (2023): 1048