INTI JOURNAL | Vol.2023:72
eISSN:2600-7320

Clustering Based on Customers' Behaviour in Accepting Personal Loan using Unsupervised Machine Learning

Lim Wai Ping[1], Goh Ching Pang[1*]

[1] Tunku Abdul Rahman University of Management and Technology Kuala Lumpur, Malaysia

[*]**Email:** gohcp@tarc.edu.my

Abstract

This research explores the application of unsupervised learning, a subset of Artificial Intelligence (AI), to analyze customer behavior in accepting personal loans within the banking sector. Focusing on clustering algorithms, the study employs popular methods like K-Means Clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Agglomerative Hierarchical Clustering, and Mean Shift Clustering to understand customer characteristics and behaviors. Using a dataset from Kaggle comprising 13 attributes and 5000 rows of bank customer data, the research addresses the challenge of processing overwhelming customer information by leveraging machine learning models. The objective is to enhance target marketing campaigns, increase success ratios, and identify potential customers with a higher probability of loan acceptance. This research contributes novel insights into the application of clustering algorithms in banking, proposing pragmatic solutions for efficient data analysis and campaign optimization. The findings underscore the pivotal role of AI in navigating and unraveling customer behavior complexities in the banking industry. All clustering models are developed successfully in producing cluster results. Besides, this system was able to perform good data visualization in order to provide better user experience. All the models are compared and discussed based on the results obtained. As a conclusion, K-Means clustering was presenting better cluster results using this particular dataset.

Keywords

K-Means Clustering, DBSCAN, Agglomerative Hierarchical Clustering, Mean Shift Clustering, Unsupervised Machine Learning

Introduction

In the contemporary landscape of rapid digital transformation, the escalating value of data has positioned it as one of the world's most indispensable resources. This paradigm shift has necessitated businesses to prioritize the collection of customer data. The advent of mature and advanced technology facilitates the capture and analysis of customer data through various channels, including browsing histories, cookies, company records, social media, and third-party trackers (Choi et al., 2019). However, the sheer volume of collected data poses a significant challenge for manual processing. Here, machine learning algorithms emerge as indispensable tools for efficiently handling vast datasets and transforming them into actionable information. Artificial

Submission: 14 November 2023; **Acceptance:** 5 December 2023

Intelligence (AI) programs equipped with sophisticated algorithms play a pivotal role by identifying anomalies and providing contextualized data-driven recommendations to organizational decision-makers (Scheidt & Staudt, 2024; Sarker, 2021; Jones & Sah., 2023; Seema et al., 2021).

The increasing integration of machine learning in business operations is driven by several compelling reasons. Foremost, it empowers sales teams to comprehensively understand the overall target market, enhancing customer experience across various aspects of a company's online presence and brand identity. Machine learning (Topuz & Çakici, 2023) not only aids in understanding customer demands and habits but also contributes to strategic decision-making in product selection, inventory expansion, and service offerings, directly impacting the bottom line of businesses. Moreover, companies that proactively track and analyze customer data gain a competitive edge, enabling them to pivot successfully in response to evolving trends or external industry influences beyond their control. The universal importance of customer data for targeted marketing, improved customer service, and informed decision-making is evident across diverse industries (Gaczek et al., 2023). In the banking sector, the possession of accurate and comprehensive customer data, encompassing details like monthly salary, daily expenses, and transaction histories, proves instrumental in devising strategic planning and promotions to engage potential customers (Alnaser et al., 2023). The integration of AI technology not only streamlines data processing but also empowers businesses to make predictive forecasts for future plans.

This research addresses the critical need to analyze and understand customer behavior patterns in the banking industry. By applying unsupervised machine learning techniques, this research aims to uncover valuable insights into customer behavior related to accepting personal loans. These insights have the potential to inform strategic decision-making, product offerings, and marketing strategies for financial institutions. This study contributes to the growing body of knowledge in the field of customer data analysis and its application in the banking sector, ultimately benefiting both businesses and customers alike.

Methodology

The dataset utilized in this research is sourced from Kaggle, a renowned platform for datasets, named "Bank_Personal_Loan_Modelling.csv." Comprising 14 attributes and 5000 rows, the dataset weighs 350.74 kB. The features and its description are listed in Table 1. Data cleaning precedes algorithmic application, a critical step ensuring dataset integrity. Employing pandas, the dataset undergoes scrutiny, with attribute names refined for clarity, missing values addressed, and extraneous features removed, rendering the dataset primed for analysis. Four clustering algorithms—K-Means Clustering, DBSCAN, Agglomerative Hierarchical Clustering, and Mean Shift Clustering—are selected for customer characteristic analysis. Specific algorithmic requirements, such as cluster number and parameters, are elucidated. The K-Means Clustering algorithm's optimal k value is determined through an elbow graph, revealing k = 4. The selected features—‘AnnIncome,’ ‘HouseMortgage,’ and ‘AnnAvgCCSpent’—guide the clustering process, assigning labels (0, 1, 2, or 3) based on the calculated k value. For DBSCAN, eps (12.5) and minimum sample (6) are determined using heatmaps, resulting in 3 clusters and 1 outlier. Agglomerative Hierarchical Clustering's k value (2) is identified through the silhouette score elbow method. Lastly, Mean Shift Clustering automatically determines 2 clusters. The systematic application of these algorithms ensures comprehensive customer behavior analysis for personal loan acceptance.

Table 1. Description of dataset

Attributes	Descriptions
ID	Customer ID
Age	Customer’s age in completed years
Experience	Years of professional experience
Income	Annual income of the customer ($000)
ZIP Code	Home Address ZIP code
Family	Family size of the customer
CCAvg	Average spending on credit cards per month ($000)
Education	Education Level. 1: Undergrad; 2: Graduate; 3: Advanced/Professional
Mortgage	Value of house mortgage if any. ($000)
Securities Account	Does the customer have a securities account with the bank?
CD Account	Does the customer have a certificate of deposit (CD) account with the bank?
Online	Does the customer use internet banking facilities?
CreditCard	Does the customer use a credit card issued by UniversalBank?
Personal Loan	Did this customer accept the personal loan offered in the last campaign?

Results and Discussion

Figure 1 depicts the outcomes derived from the K-Means Clustering model. The left diagram offers a perspective utilizing ‘HouseMortgage’ and ‘AnnIncome’ as axes, while the right diagram employs ‘HouseMortgage’ and ‘AnnAvgCCSpent.’ Each cluster is distinctly labeled for clarity. Upon scrutinizing both diagrams, the clusters emerge as meaningful and interpretable, each possessing unique characteristics. Cluster 0 encapsulates customers with lower annual income and annual average credit card spending, devoid of house mortgages. In contrast, cluster 1 represents customers burdened with higher house mortgages. Cluster 2 is characterized by customers with lower house mortgages, while cluster 3 denotes customers with higher annual income and annual average credit card spending, without house mortgages. This visual representation enhances the understanding of customer segmentation based on critical attributes, facilitating nuanced insights into distinct customer profiles.

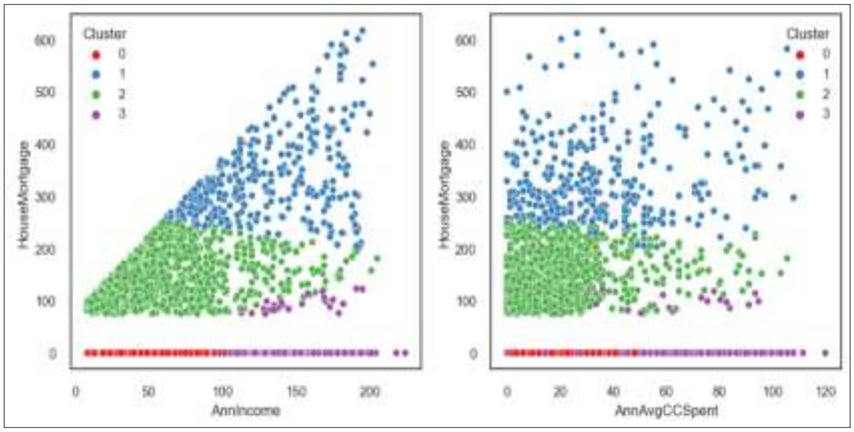

Figure 1. Result of K-Means Clustering model

Figure 2 illustrates the outcomes of the DBSCAN analysis through two distinct diagrams. The left diagram, utilizing ‘HouseMortgage’ and ‘AnnIncome’ as axes, presents a graphical representation of the results. On the right, the diagram incorporates ‘HouseMortgage’ and ‘AnnAvgCCSpent’ to offer an alternative perspective. Clear labels are provided for each cluster, enhancing comprehension. In detail, cluster 0 identifies customers without a house mortgage, irrespective of their annual income and annual average credit card spending. Cluster 1 represents customers with lower annual income and reduced annual average credit card spending. Lastly, cluster 2 characterizes customers with moderate annual income and diminished annual average credit card spending. It's notable that numerous data samples remain unclustered, being treated as outliers in the analysis. This comprehensive depiction aids in understanding the nuanced segmentation of customers based on diverse attributes.

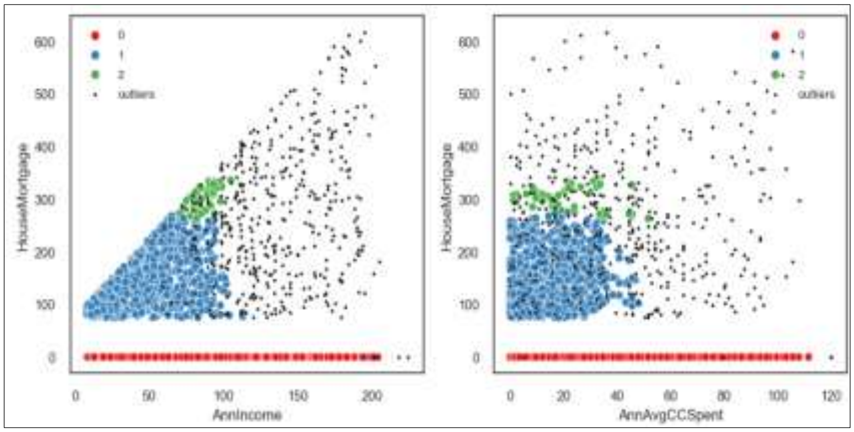

Figure 2. Result for DBSCAN model

Figure 3 presents the outcomes of the Agglomerative Hierarchical Clustering in two visual representations. The left diagram, employing 'HouseMortgage' and 'AnnIncome' as axes, showcases the clustering based on this perspective. Simultaneously, the right diagram utilizes 'HouseMortgage' and 'AnnAvgCCSpent' to offer an alternative view of the clustering results. Clear and informative labels are incorporated for each cluster, enhancing the interpretability of the diagrams. In-depth analysis reveals that cluster 0 comprises customers with higher house mortgages in relation to their annual income and annual average credit card spending. On the other hand, cluster 1 represents customers with lower house mortgages considering their annual income and annual average credit card spending. This visualization provides valuable insights into the segmentation of customers based on distinct financial attributes, aiding in the interpretation of clustering patterns and their implications.

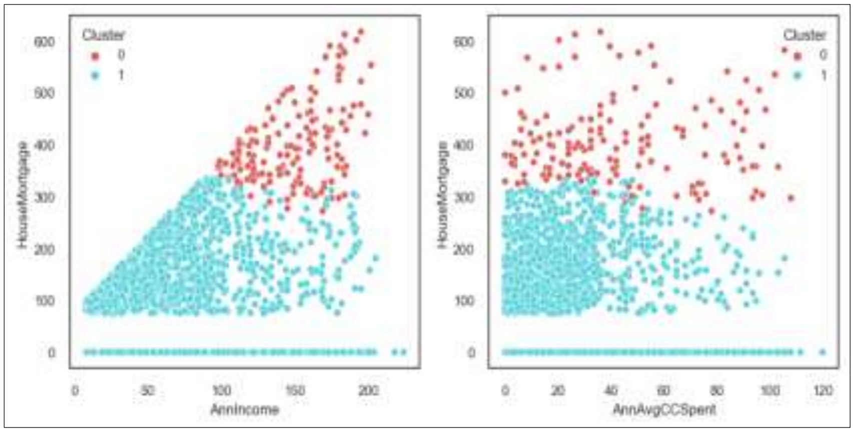

Figure 3. Result for Agglomerative Hierarchical Clustering model

Figure 4 exhibits the outcomes of Mean Shift Clustering through two insightful diagrams. The left diagram, visualizing 'HouseMortgage' and 'AnnIncome,' elucidates the clustering results from this perspective. Simultaneously, the right diagram, employing 'HouseMortgage' and 'AnnAvgCCSpent,' offers an additional viewpoint. The inclusion of descriptive labels for each cluster enhances the comprehensibility of the diagrams. In-depth analysis of the results reveals that cluster 0 encompasses customers with house mortgages in relation to their annual income and annual average credit card spending. Conversely, cluster 1 represents customers without house mortgages, considering both their annual income and annual average credit card spending. This visualization aids in the interpretation of Mean Shift Clustering patterns, providing valuable insights into the segmentation of customers based on their financial attributes.

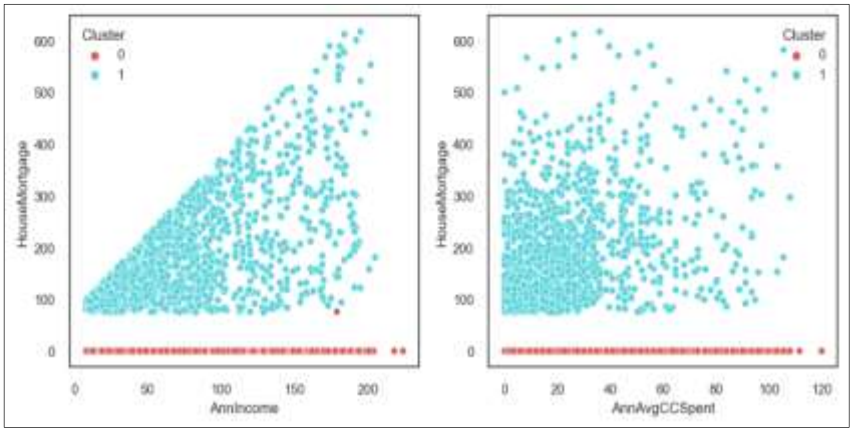

Figure 4. Result for Mean Shift Clustering model

Conclusion

The K-Means clustering model emerges as the most compelling and suitable choice for clustering the given dataset. This model delineates four distinct clusters with reasonably balanced distribution sizes, effectively capturing customer behavior related to accepting personal loans. In contrast, the DBSCAN model faces challenges due to its reliance on two critical parameters, eps and minimum sample values, potentially excluding many data samples from clustering. This limitation results in a significant loss of valuable data samples. The Agglomerative Hierarchical Clustering and Mean Shift Clustering models encounter a similar issue—oversimplification in clustering. The selected features do not achieve optimal utilization in clustering, leading to less convincing results. Despite these challenges, all clustering models successfully produce cluster results, and the system excels in data visualization, enhancing user experience. In summary, K-Means clustering outperforms other models in delivering superior cluster results for this specific dataset. However, there are several promising directions for further research and development in this field: a) refinement of clustering algorithms: future studies can focus on refining existing clustering algorithms or developing new ones that are specifically tailored to the nuances of customer behavior in the banking industry. This may involve addressing the limitations of DBSCAN and other models to improve their performance and b) feature engineering: investigate the possibility of identifying additional relevant features or fine-tuning the feature selection process to enhance the accuracy of clustering models. Feature engineering plays a crucial role in the success of machine learning algorithms.

Acknowledgements

This research is supported by Tunku Abdul Rahman University of Management and Technology, Malaysia.

References

Alnaser, F., Rahi, S., Alghizzawi, M., & Ngah, A. H. (2023). Does artificial intelligence (AI) boost digital banking user satisfaction? Integration of expectation confirmation model and antecedents of artificial intelligence enabled digital banking. *Heliyon, 9*(8), e18930. https://doi.org/10.1016/j.heliyon.2023.e18930

Choi, J. P., Jeon, D.-S., & Kim, B.-C. (2019). Privacy and personal data collection with information externalities. *Journal of Public Economics, 173*, 113–124. https://doi.org/10.1016/j.jpubeco.2019.02.001

Gaczek, P., Leszczyński, G., & Mouakher, A. (2023). Collaboration with machines in B2B marketing: Overcoming managers' aversion to AI-CRM with explainability. *Industrial Marketing Management*, *115*, 127–142. https://doi.org/10.1016/j.indmarman.2023.09.007

Jones, K., & Sah, S. (2023). The implementation of machine learning in the insurance industry with big data analytics. *International Journal of Data Informatics and Intelligent Computing*, *2*(2), 21–38. https://doi.org/10.59461/ijdiic.v2i2.47

Rawat, S., Rawat, A., Kumar, D., & Sabitha, A. S. (2021). Application of machine learning and data visualization techniques for decision support in the insurance sector. *International*

Journal of Information Management Data Insights, *1*(2), 100012. https://doi.org/10.1016/j.jjimei.2021.100012

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, *2*, 160. https://doi.org/10.1007/s42979-021-00592-x

Scheidt, F., & Staudt, P. (2024). A data-driven recommendation tool for sustainable utility service bundles. *Applied Energy*, *353*, 122137. https://doi.org/10.1016/j.apenergy.2023.122137

Topuz, B., & Çakici Alp, N. (2023). Machine learning in architecture. *Automation in Construction*, *154*, 105012. https://doi.org/10.1016/j.autcon.2023.105012