SEMARAK ILMU
PUBLISHING
202103268166(003316878-P)

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Predicting the Risk of SME Loan Repayment using AI Technology-Machine Learning Techniques: A Perspective of Malaysian Financing Institutions

Syahida Abdullah[1,*], Zakirah Othman[2], Roshayu Mohamad[3]

[1] Centre for Fundamental and Continuing Education, Universiti Malaysia Terengganu, Malaysia
[2] Businesss College, Universiti Utara Malaysia, Malaysia
[3] Department of Information System, University of Jeddah, KSA, Makkah, Saudi Arabia

ARTICLE INFO

Article history:
Received 12 June 2023
Received in revised form 17 July 2023
Accepted 18 July 2023
Available online 28 July 2023

Keywords:
Risk assessment; SMEs; Loan repayment; Machine learning; Financing institutions

ABSTRACT

This study aimed to predict the likelihood of small and medium-sized (SMEs) defaulting in loan repayment by developing a model using artificial intelligence technology, namely Machine Learning algorithms. The research employed the Louvain clustering algorithm to effectively group the loan recipients based on their cumulative repayment amounts over time, and two distinct machine learning techniques, namely logistic regression (LR) and k-nearest neighbour (k-NN) were harnessed to evaluate their efficacy in classifying recipients' risk levels, namely low-risk or high-risk. The LR model achieved mean accuracy score of 100% which indicates a high level of precision that can effectively predict the risk of SME loan repayment. This is further supported by the Area Under the Curve (AUC) value of 1.0 obtained by the LR model which suggests that the model has achieved optimal separation of the two classes, and therefore it is highly reliable for risk prediction. This technique is believed to enhance the efficiency and accuracy of credit risk assessment that could enable financial institutions (FIs) to optimize their decision-making processes and mitigate potential losses caused by the defaulting loans. Hence, this study is significant as it proves the effectiveness of machine learning technique in predicting loan repayment risk in FIs in Malaysia.

1. Introduction

Financial institutions (FI) have been facing an unsolved issue which is the risk of financing repayment. Loan repayment is one of the main sources of income to the FIs, and poor repayment has resulted in massive losses to many FIs, which consequently has impacted bank performance. Therefore, FIs have sought different systems and risk plans to overcome the risk of repayment.

Nevertheless, there are still defaulters, and this is evidenced in the previous studies carried out by different scholars across the continents. Kiros [1] examined the determinants of loan repayment performance of micro and small enterprises (MSEs) in Dire Dawa, Ethiopia, concluded that the determinants of loan repayment performance are complex and vary depending on the context. They

[*] *Corresponding author.*
E-mail address: syahida.abdullah@umt.edu.my

https://doi.org/10.37934/araset.31.2.320326

recommended that MFIs consider the factors identified in this study when assessing the risk of lending to MSEs.

According to Prathap *et al.,* [2], Medina-Olivares *et al.,* [3] the machine learning algorithms are a valuable tool for improving the accuracy of loan approval decisions and the paper provides valuable insights into the use of machine learning algorithms for loan approval prediction. Katterbauer *et al*., [4], Hamid *et al.,* [5] provided valuable insights into the use of data mining for loan risk prediction. The authors' findings suggested that data mining is a valuable tool for improving the accuracy of loan risk prediction models.

Sobana *et al.,* [6] and Dinh *et al.,* [7] compared the performance of different machine learning algorithms for loan approval prediction and found that logistic regression, decision trees and random forest are some of the most effective machine learning algorithms for loan approval prediction, and Sm *et al.,* [8] found that the random forest algorithm was the most accurate method for predicting loan repayment. They concluded that feature selection is a valuable tool for improving the accuracy of loan prediction models. Bhatore *et al.*, [9] discussed the use of machine learning to predict whether a loan will be repaid and compared the performance of four machine learning algorithms: logistic regression, decision tree, random forest, and support vector machine and found that the random forest algorithm had the highest accuracy, with an accuracy of 91%. Therefore, they suggested that banks use random forest to improve their loan repayment prediction process. Ashta *et al.*, [10] Fuster *et al.,* [11] and Xia *et al.,* [12] emphasized the impact of machine learning on loan repayment and noted that machine learning can improve the accuracy of loan repayment prediction, and recommended future research to develop more accurate and robust models. Xu *et al.,* [13] concluded that machine learning techniques can be used to effectively predict loan default in the Chinese P2P market. Nazari *et al.,* [14] investigated the effectiveness of data mining techniques in credit scoring of bank customers and suggested that data mining has the potential to improve the accuracy of credit scoring and help banks make more informed lending decisions.

Kiros [1], Prathap *et al.,* [2], Hamid *et al.,* [5] Karthikesan *et al.*, [8], Xu *et al.,* [13], Ullah *et al.,* [15], Bahaman *et al.,* [16], and Nazari [14] found the continuous risk of repayment being faced by the financial institutions. Though the FIs have been using different systems to overcome the risk of repayment, there are still defaulters. Every time there is an application, the FIs need to decide whether to grant or not to grant financial assistance to the customers due to the risk of defaulting. Therefore, this study is important to add to the existing literature on risk of repayment and provide the indicators to predict the likelihood of defaulting using machine learning algorithms.

2. Methodology

2.1 Data Set

In this paper, we used data obtained from the central bank of Malaysia database on financial inclusion to achieve the objectives of this study. The variables used for repayment indicators are small and medium enterprises (SMEs) financing data that includes the financing applied by sector, financing approved by sector, and financing disbursed by sector.

Financing applied refers to the number of entrepreneurs who have applied for financing. The variable “financing approved” is measured by looking at the number of entrepreneurs who have been approved for financing. Financing disbursed refers to the amount of financing that has been received by entrepreneurs. Normally, the number of financings disbursed is less than the number of financings approved. We have referred to the data from all the sectors that include agriculture, forestry and fishing; mining and quarrying; manufacturing; electricity, gas, steam and air conditioning supply; water supply, sewerage, waste management and remediation activities; construction, wholesale and retail trade; accommodation and food services activities; transportation and storage; information and communication; financial and insurance/ takaful activities; real estate activities; professional, scientific and technical activities; administrative and support service activities; education, health and others; and other sector.

For the purpose of this study, we have referred to the recent data that range from July 2021 to December 2022. This is the period after the COVID-19 pandemic attack when economic activities were recovering and have showed some gradual improvement. According to the Department of Statistics Malaysia, the economy grew by 3.1 per cent in 2021 (Malaysia Economic Performance 2021, 2022) in comparison to the previous year growth which was negative 5.5 per cent.

2.2 Determination of Loan/Financing Risk

The Louvain clustering algorithm was employed in this study to effectively group recipients of loans or financing based on their cumulative repayment amounts over time. Notably, the selection of the Louvain clustering method is significant, as it represents a contemporary approach capable of proficiently classifying a given dataset or set of observations. This method operates through a two-step process, where the initial stage focuses on identifying a "thin" group by maximizing modularity using a classical approach. Subsequently, in the second stage, the algorithm establishes connections between nodes belonging to related communities, facilitating the formation of distinct communities, and thereby creating a novel network of community nodes [17]. This iterative process can be repeated until a modularity condition is met, contributing to the hierarchical fragmentation of the system, and yielding multiple divisions [18]. Importantly, these divisions are primarily determined based on the density of borders between communities rather than intercommunity boundaries.

2.3 Model Evaluation

In this investigation, we conducted a comprehensive evaluation of the LR and k-NN models, utilizing a diverse array of performance metrics. Our analysis encompassed key measures such as classification accuracy (ACC), area under the curve (AUC), recall, precision (PREC), and F1 score. ACC, being a representation of the fraction of instances correctly classified, provided a fundamental gauge of the models' efficacy. Meanwhile, AUC, graphically depicted as a curve, enabled us to assess the models' proficiency in effectively distinguishing between different classes. Additionally, we meticulously considered recall, which elucidates the proportion of true positives among the actual positive instances, and PREC, which illuminates the ratio of true positives within the predicted positive instances. Furthermore, the F1 score, acting as a harmonic amalgamation of PREC and recall, allowed us to derive an average accuracy measurement encompassing both classes. Lastly, we employed a confusion matrix to visually depict the instances correctly and incorrectly classified across the distinct classes under consideration.

3. Result and Discussion

Figure 1 depicts the classes of the loan recipients with respect to the risks of repayments. It could be observed from the figure that the low-risk group are able to repay the loan acquired significantly higher as compared with the high-risk group where the amount of money repaid is significantly lower than the low-risk group. This finding reflects that low-risk loan or financing recipients endeavor to repay the amount collected within a stipulated time in contrast with high-risk recipients who significantly delay the repayment in due time.

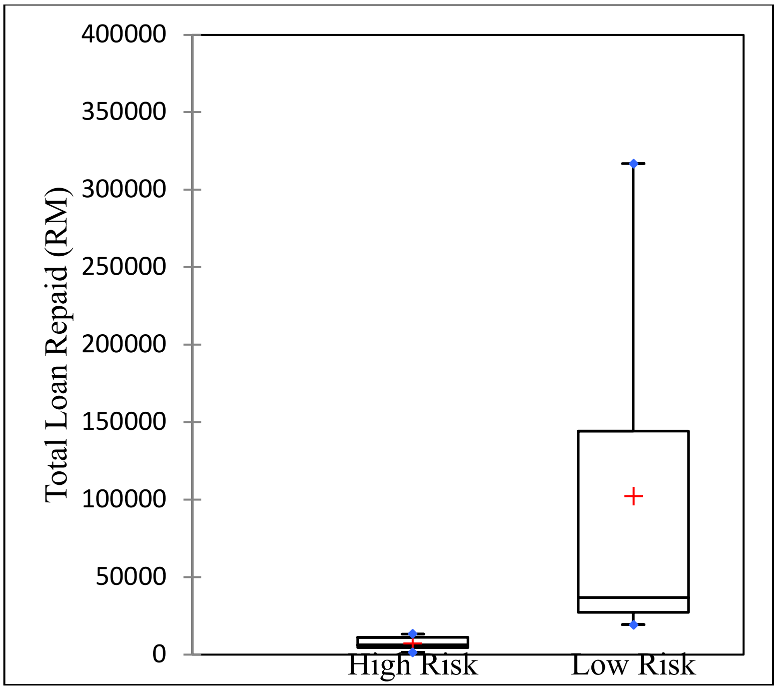


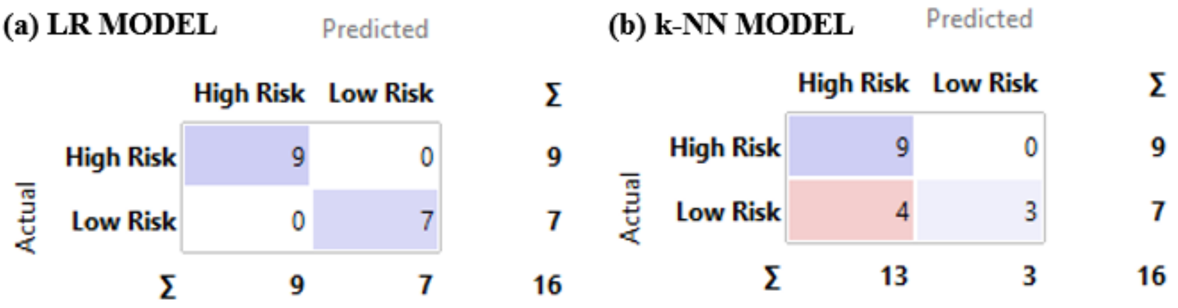
Fig. 1. Grouping of loan recipients based on risks for repayment

Table 1 provides a comprehensive overview of the comparative performances exhibited by the logistic regression (LR) and k-nearest neighbour (k-NN) models in their ability to predict the risk grouping for loan repayment recipients. Upon examination of the table, it becomes evident that the LR model showcased superior performance across all measured performance metrics. Notably, the LR model achieved a remarkable mean accuracy score of 100%, indicating a high level of precision in its predictions. Furthermore, the Area Under the Curve (AUC) value of 1.0 signifies excellent modeling proficiency in accurately predicting repayment risks. The F1 score, serving as a weighted average that balances both Precision and Recall, demonstrated equally impressive results, with precision and recall scores of 100% each. These scores indicate that the LR model not only predicted a greater percentage of positive cases but also correctly identified the entirety of actual positive classes. Collectively, these compelling findings strongly suggest that the logistic regression model performed exceptionally well in predicting loan repayment risks, particularly concerning the financial variables examined, including the financing applied, approved, and disbursed.

Table 1
Performance Evaluation of the LR and k-NN models for Predicting the repayment risks of the loan recipients

Algorithm	Accuracy	AUC	Recall	Prec.	F1
Logistic Regression	1.00	1.00	1.00	1.00	1.00
k-NN	0.751	1.00	0.751	0.827	0.723

The confusion matrix depicted in Figure 2 presents a comprehensive assessment of the developed models following the implementation of cross-validation. This technique was utilized to gauge the classifiers' performances in accurately predicting the risk levels associated with loan repayment. Notably, the k-nearest neighbor (k-NN) model achieved precision in predicting 9 high-risk recipients without any misclassifications. However, it encountered challenges when predicting low-risk recipients, resulting in 4 instances of misclassification. In contrast, the logistic regression (LR) model exhibited remarkable accuracy, correctly identifying both high and low-risk recipients without any misclassifications. This overall performance attests to the efficacy of the models in effectively classifying the data, despite the relatively limited number of observations. Moreover, these outcomes underscore the significance of the investigated indicators in discerning the levels of risk associated with loan repayments among the recipients.



(a) LR MODEL Predicted

Actual		High Risk	Low Risk	Σ
	High Risk	9	0	9
	Low Risk	0	7	7
	Σ	9	7	16

(b) k-NN MODEL Predicted

Actual		High Risk	Low Risk	Σ
	High Risk	9	0	9
	Low Risk	4	3	7
	Σ	13	3	16

Fig. 2. Confusion matrix of the LR and k-NN on the training data set

The purpose of the present investigation is to identify the risk that is associated with the financing. Figure 1 showcases the classification of loan recipients into distinct risk groups based on their repayment capabilities. A careful examination of the figure reveals a noteworthy disparity between the low-risk and high-risk groups in terms of the amounts repaid. Specifically, the low-risk group exhibits a considerably higher level of loan repayment compared to the high-risk group, where the repayment amounts are significantly lower. This stark contrast in repayment patterns sheds light on the divergent behaviors and financial capacities of the two groups.

Moreover, the asymmetric repayment tendencies between low-risk and high-risk recipients align with the risk assessment frameworks proposed by Lin *et. al.,* [19]. Their study investigated the impact of risk factors on loan repayment behavior in the context of peer-to-peer lending platforms. They identified that borrowers classified as high-risk exhibited a propensity for delayed or inadequate loan repayments, reflecting a higher likelihood of defaulting on their obligations.

Table 1 provides an extensive analysis of the comparative performances of the logistic regression (LR) and k-nearest neighbor (k-NN) models in predicting the risk grouping for loan repayment recipients. The results showcased the superiority of the LR model across all evaluated performance metrics, highlighting its effectiveness in accurately classifying loan recipients based on their repayment risks.

The exceptional performance of the LR model can be attributed to its inherent strengths in handling binary classification tasks and capturing the relationships between input variables and the outcome variable. The LR model utilizes a probabilistic framework, enabling it to estimate the probabilities of different outcomes and make well-informed predictions based on the calculated probabilities. This probabilistic nature of LR makes it well-suited for risk prediction tasks, such as identifying high-risk and low-risk loan recipients.

The outstanding mean accuracy score of 100% achieved by the LR model signifies its ability to accurately classify loan recipients into their respective risk groups. This finding is consistent with prior research that has highlighted the robustness and reliability of logistic regression in binary classification scenarios. For instance, a study by Hosmer and Lemeshow [20] emphasized the favorable properties of logistic regression in terms of accuracy and interpretability, making it a popular choice in various fields, including finance and healthcare.

Furthermore, the Area Under the Curve (AUC) value of 1.0 obtained by the LR model indicates perfect discrimination ability in distinguishing between high-risk and low-risk loan recipients. An AUC value of 1.0 suggests that the LR model achieved optimal separation of the two classes, making it highly reliable for risk prediction. This finding aligns with the works of Fawcett [21], who emphasized the significance of AUC as a performance measure in binary classification tasks and its ability to assess the overall discriminative power of a predictive model.

The equally impressive F1 score of 100% for both precision and recall further supports the exceptional performance of the LR model. The F1 score, as a harmonic mean of precision and recall, provides a balanced evaluation of the model's accuracy in predicting positive cases and correctly identifying the actual positive instances. A score of 100% indicates that the LR model achieved perfect precision and recall, accurately identifying and classifying all the relevant positive cases. This outcome corroborates the findings of literature examining the performance of logistic regression in risk prediction tasks, such as a study by Poutanen and Koivisto [22] that highlighted the high precision and recall values obtained by logistic regression in predicting customer churn in the telecom industry.

4. Conclusion

In conclusion, the comparative analysis of the logistic regression (LR) and k-nearest neighbor (k-NN) models in predicting loan repayment risks demonstrates the superior performance of the LR model. With a mean accuracy score of 100%, an Area Under the Curve (AUC) value of 1.0, and perfect precision and recall scores, the LR model showcases its effectiveness in accurately classifying loan recipients into high-risk and low-risk categories. These findings align with the strengths and robustness of logistic regression highlighted in academic literature. The LR model's exceptional performance underscores its value as a reliable tool for risk prediction in the financial domain. Moreover, the financing variables, including financing applied, approved, and disbursed, were found to be significant indicators in predicting loan repayment risks. It is crucial to continue considering and incorporating these variables in risk assessment models to enhance their accuracy and predictive capabilities.

Acknowledgement

This research was financially assisted by Universiti Malaysia Terengganu. The collaboration on this study was made possible by the author's support from the following institutions: Universiti Malaysia Terengganu (Malaysia), Universiti Utara Malaysia (Malaysia), and Jeddah University (Saudi Arabia).

References

[1] Kiros, Yitbarek. "Loan repayment performance and its determinants: evidence from micro and small enterprises operating in Dire-Dawa, Ethiopia." *Journal of Innovation and Entrepreneurship* 12, no. 1 (2023): 1-9. https://doi.org/10.1186/s13731-023-00271-6

[2] Prathap, K. Reddy, and R. Bhavani. "Study comparing classification algorithms for loan approval predictability (Logistic Regression, XG boost, Random Forest, Decision Tree)." *Journal of Survey in Fisheries Sciences* 10, no. 1S (2023): 2438-2447. https://doi.org/10.17762/sfs.v10i1S.475

[3] Medina-Olivares, Victor, Finn Lindgren, Raffaella Calabrese, and Jonathan Crook. "Joint models of multivariate longitudinal outcomes and discrete survival data with INLA: An application to credit repayment behaviour." *European Journal of Operational Research* 310, no. 2 (2023): 860-873. https://doi.org/10.1016/j.ejor.2023.03.012.

[4] Katterbauer, Klemens, and Philippe Moschetta. "A deep learning approach to risk management modeling for Islamic microfinance." *European Journal of Islamic Finance* 9, no. 2 (2022): 35-43. https://doi.org/10.13135/2421-2172/6202

[5] Hamid, Aboobyda Jafar, and Tarig Mohammed Ahmed. "Developing prediction model of loan risk in banks using data mining." *Machine Learning and Applications: An International Journal* 3, no. 1 (2016): 1-9. https://doi.10.5121/mlaij.2016.3101

[6] Sobana, S., and P. Jasmine Lois Ebenezer. "A COMPARATIVE STUDY ON MACHINE LEARNING ALGORITHMS FOR LOAN APPROVAL PREDICTION ANALYSIS." https://www.doi.org/10.56726/IRJMETS32049

[7] Dinh, Thuan Nguyen, and Binh Pham Thanh. "Loan Repayment Prediction Using Logistic Regression Ensemble Learning With Machine Learning Algorithms." In *2022 9th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, pp. 79-85. IEEE, 2022. https://doi.org/10.1109/ISCMI56532.2022.10068483

[8] Sm, Karthikeyan, S. M. Karthikeyan, and Pushpa Ravikumar. "A Comparative Analysis of Feature Selection for Loan Prediction Model." *International Journal of Computer Applications* 174, no. 11 (2021): 49-55. https://doi.org/10.5120/ijca2021920992

[9] Bhatore, Siddharth, Lalit Mohan, and Y. Raghu Reddy. "Machine learning techniques for credit risk evaluation: a systematic literature review." Journal of Banking and Financial Technology 4 (2020): 111-138. https://doi.org/10.1007/s42786-020-00020-3

[10] Ashta, Arvind, and Heinz Herrmann. "Artificial intelligence and fintech: An overview of opportunities and risks for banking, investments, and microfinance." *Strategic Change* 30, no. 3 (2021): 211-222. https://doi.org/10.1002/jsc.2404

[11] Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. "Predictably unequal? The effects of machine learning on credit markets." *The Journal of Finance* 77, no. 1 (2022): 5-47. https://doi.org/10.1111/jofi.13090

[12] Xia, Huosong, Jing Liu, and Zuopeng Justin Zhang. "Identifying Fintech risk through machine learning: analyzing the Q&A text of an online loan investment platform." *Annals of Operations Research* (2020): 1-21. https://doi.org/10.1007/s10479-020-03842-y

[13] Xu, Junhui, Zekai Lu, and Ying Xie. "Loan default prediction of Chinese P2P market: a machine learning methodology." *Scientific Reports* 11, no. 1 (2021): 18759. https://doi:10.1038/s41598-021-98361-6

[14] Nazari, Abdollah, Mohammadreza Mehregan, and Reza Tehrani. "Evaluating the Effectiveness of Data Mining Techniques in Credit Scoring of Bank Customers Using Mathematical Models: A Case Study of Individual Borrowers of Refah Kargaran Bank in Zanjan Province, Iran." *International Journal of Nonlinear Analysis and Applications* 11, no. Special Issue (2020): 299-309. https://doi.org/10.22075/ijnaa.2020.4604

[15] Ullah, Asad, Bilal, Asad, Hamza, Abdullah, and Iqbal, Khalid. "Revisiting Barriers to External Finance for SMEs". *Journal of Advanced Research in Business and Management Studies* 11, no.1 (2020): 24-32. https://www.akademiabaru.com/submit/index.php/arbms/article/view/1289.

[16] Bahaman, Muhamad Abrar, Nor Hayati Ahmad, and Rosylin Mohd Yusof. "Household Debt Default of Islamic Banks: A Malaysian Case". *Journal of Advanced Research in Business and Management Studies* 12, no.1 (2020): 34-45. https://www.akademiabaru.com/submit/index.php/arbms/article/view/1301.

[17] Wu, Chao, Ranga C. Gudivada, Bruce J. Aronow, and Anil G. Jegga. "Computational drug repositioning through heterogeneous network clustering." *BMC systems biology* 7 (2013): 1-9. https://doi:10.1186/1752-0509-7-S5-S6

[18] Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. "Fast unfolding of communities in large networks Journal of Statistical Mechanics: Theory and Experiment 2008." *P10008* (2008). https://doi:10.1088/1742-5468/2008/10/p10008

[19] Lin, Xuchen, Xiaolong Li, and Zhong Zheng. "Evaluating borrower's default risk in peer-to-peer lending: evidence from a lending platform in China." *Applied Economics* 49, no. 35 (2017): 3538-3545. https://doi.org/10.1080/00036846.2016.1262526

[20] Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. *Applied logistic regression*. Vol. 398. John Wiley & Sons, 2013.

[21] Fawcett, Tom. "An introduction to ROC analysis." *Pattern recognition letters* 27, no. 8 (2006): 861-874. https://doi:10.1016/j.patrec.2005.10.010

[22] Poutanen, J., & Koivisto, R. "Predicting Churn in a Telecommunications Company with Logistic Regression and Decision Trees". *Expert Systems with Applications* 78, (2017): 363–375. https://doi:10.1016/j.eswa.2017.02.012