Applied Mathematics and Computational Intelligence
Volume 12, No. 2, Aug 2023 [36-47]

Prediction on Loan Defaults: Tree-based Approach

Balqis Adzman[1], Sayang Mohd Deni[2]*, Mohamad Ismeth Khan Azhar Suhaimi[2]

[1]Level 10, Tower RHB Centre, Jalan Tun Razak, 50400 Kuala Lumpur, Malaysia.
[2]School of Mathematical Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA (UiTM) Shah Alam, 40450 Selangor, Malaysia.

* Corresponding author : sayan929@uitm.edu.my

Received: 27 Oct 2022
Accepted: 14 July 2023

ABSTRACT

Financial institutions have been exploring the application of machine learning approaches due to its exceptional performance as well as overwhelming exposure, especially in predicting the repayment ability of their customers. The ability of machine learning methods in dealing with big and more complicated data structure has made it favourable as the financial data is often very large and complex in nature. Thus, this study adopts two tree-based machine learning approaches to predict the loan defaulters, namely random forest and extreme gradient boosting (XGBoost). However, due to its sensitivity towards imbalanced dataset, this study has addressed this issue beforehand. The performance of both approaches was assessed by computing the accuracy, precision, recall, F-1 score, ROC as well as AUC. XGBoost proves to be able to outperform the traditional machine learning model, random forest, with 61.77% accuracy, other than it generally takes lower computation time. The model is able to report higher value for all the assessment matrices used. Other than that, this study also focuses on the customers' demographic information and found that it was useful in predicting their repayment ability especially the customer's length of service, education level as well as age.

Keywords: Loan Defaults Prediction, XGBoost, Random Forest, Imbalanced Dataset, Machine Learning Model

1 INTRODUCTION

Defaulted loan or non-performing loan (NPL) is defined as a loan in which the borrower has failed to make repayments as per the payment scheduled by the bank. Defaulted loan is huge threat to the banks as it can lead to the loss that the bank will experience and often indicates the negligence of the bank in managing its credit risk policy. In Malaysia, it is recorded that defaulted loan trend increased moderately in 2008 but has shown a downward trend from that point onwards until 2016 [1]. However, the current Covid-19 pandemic has affected the global economy as well as becoming a contributing factor to the rising of defaulted loans. Due to the impact may causes many to lose their job and source of income which it is hard for them to meet their obligation of paying their debts or loans. Based on previous studies have shown that the ability of the customers to repay their financial obligation can be explained by the macroeconomic factors as well as the information on customers' loan such as their liability and payment behaviour. The machine learning methods have capability in handling big data due to the high dimensional nature of the customers data. For instance, other applications of machine learning approach on predictive models could be found in various field of studies such as intervention and prevention of diabetic retinopathy [2], diagnosis of Alzheimer's disease [3], prediction of dengue outbreaks [4], prediction and classification on future PM10 concentrations [5] and rainfall prediction in flood prone areas [6]. However, it is worth to highlight the inherent characteristic of loan data, especially when it tends to be imbalanced and need to be treated to avoid misleading prediction outcome.

Aforementioned, the loan defaults can be classified into two categories such as if customers are obliged repayment on their credit facility within 90 days or more. For prediction model, XGBoost has been proposed by [7] due to its performance in determining the loan defaults compared to random forest. However, for imbalanced issued, random forest has been used in minimizing the time complexity of the method and it was found outperformed against other algorithms ([8], [9]). Furthermore, XGBoost is an improved algorithm which has been used due to the performance of best time complexities and the method was acceptable to be compared from the accuracy perspective [10]. The gap of the study is due to a more complicated data structure of customer's loan, logistics regression is not feasible to adopt due to its linearity assumption between independent and dependent variables. Other than that, imbalanced data structure may produce a bias in classification model because logistics regression did not address this issue. Hence, for the purpose of comparison and alternative solution, the focus will be given only to random forest and XGBoost as well as to close the gap because both methods have their own ability to output more accurate prediction. Apart from that, this study aims to resolve the imbalanced classification problem by applying down-sampling method, to compare the performance of predictive model produced between XGBoost and random forest and to determine the influential contributing factors in the default of loan repayment. Hence, this paper discusses on data and methods, results, conclusion and recommendation. This would help the structure of the paper in a good way of representation the outcome of the study.

1.1 Loan Default Definition

In the [11] guideline, it was stated that, according to Basel II framework, there are two events that help the banks to determine whether a customer is default or not. A customer will be tagged as default if either one or both events take place. The first event that indicates if a customer default is when the bank finds that he/she is unlikely to meet their loan obligations to the bank. The other event is when the customer is more than 90 days past due on any of the financial obligation. In other words, the customer has been late to make the obliged repayment on his credit facility for more than 90 days. However, the banks have the flexibility to reclassify the customer from to default to non-default category based on the banks' definite criteria. Normally, banks will implement a more stringent criteria to classify the default loans and penalize the customers that fall into those criteria.

1.2 Factors Lead to Loan Defaults

There are several factors that influenced the default of loans. According to [12], as cited in [13], these factors can be grouped into macroeconomic factors and bank-specific factors. Gross domestic product (GDP), inflation rate and the money stock were the macroeconomic factors that have been included in their study. However, from the reviews, no other researchers have considered macroeconomic factors in predicting the loan default.

Loan amount, interest rate, instalment amount and loan grade were some of the common factors that have been employed by [14], [7], [9], [13] and [15]. These factors were included in the model based on the author's general understanding of their influence on the ability of the client to repay the loan [7]. Furthermore, [14] have included outstanding principal amount, recoveries and other factors in their predicting model as well. A different approach was applied, where they grouped 96 variables into 20 clusters which was then used to train the model.

Other than the client's loan information, their socioeconomic information such as home ownership status and location were also being adopted by [7], [9], [16] and [15], has also included customer's age and found that customer's age and location were the most crucial factors in predicting whether or not the customer will default on their loan. From the review, there can be observed that very few have included the demographic profile of the customers as the factors that affect the ability of repayment such as age, education level and length of employment in the recent studies. Assuming that all customers are not exposed to the same opportunities is another way to address demographic relationship with delinquency. Hence, in this study, those features will be included to determine the customer's ability to repay their loan obligation.

1.3 Loan Defaults Prediction Model

Over the years, numerous research and studies have been conducted by adopting the application of machine learning techniques on various areas in finance and banking domain. Prediction modeling is one of the areas that has intrigued the interest of many researchers to implement the application of machine learning as opposed to traditional statistical methods. Based on the review of some of the literatures from previous studies, decision tree and random forest are amongst the most common algorithm that have been applied by previous researchers in their studies to predict the loan default. [9], in their paper have focused solely on the application of decision tree and random forest and compared the result obtained from both techniques. They concluded that random forest provided better accuracy as compared to decision tree with 80% accuracy ratio. This finding is consistent with a study conducted by [15] where they observed that random forest outperformed the decision tree with accuracy ratio of 98% and 95%, respectively.

[7] has pointed out the objective of his research of to demonstrate how machine learning techniques can be leveraged by loan approval in predicting the probability that a client will default on their loan. [7] has adopted the application of several machine learning techniques, comprised of random forest, neural network, extreme gradient boosting (XGBoost) and ensemble model. He has also included logistic regression in his study for the purpose of comparison. He concluded that XGBoost with class weights performed the best as compared to other methods with overall accuracy of 68%. The model also showed highest AUC, i.e. 0.74, for the ROC curve compared to other methods. Another study that has adopted the application of XGBoost was conducted by [17] where they have compared the performance of XGBoost with LightGBM and found that LightGBM performed better than XGBoost. [16], in her study has also adopted the application of XGBoost in predicting the bank loan default. However, she did not conduct a comparative study, hence, it cannot be concluded if XGBoost was the most appropriate model in predicting bank loan default with her dataset.

As mentioned previously, [14] have utilized four algorithms in identifying the attributes that drive towards loan defaults, delay in repayments and the criteria of customers that will honor their loan obligations. Out of the four methods (i.e. decision tree, logistic regression, neural network and random forest), they observed that HP forest was the best model due to the lowest misclassification rate that it provided. The misclassification rate produced by random forest is only 0.0562, while the misclassification rate resulted from decision tree, neural network and logistic regression are 0.0597, 0.0767 and 0.0802, respectively.

2 MATERIAL AND METHODS

This study adopted a quantitative and correlation research as it aims to establish a prediction model and determine the influential factors. This study utilized a secondary data of customer's loan application. This data was retrieved from Kaggle2 shared by Gaurav Dutta. There are 13 independent variables such as gender, age, income, level of highest education, length of service, income type, own a car, own a house, marital status, credit amount of the loan, loan annuity, credit amount of the previous loan, loan annuity of the previous loan and a dependent variable such as customer difficulties in repayment, in order to study the loan defaults with 21,556 total of observations.

2.1 Data Description

Dependent variable (customer has difficulties in repayment) will be transformed in terms of 1 and 0, which count as two classes in a variable and the rest are independent variables. Right after the data were acquired, it is vital to understand the information of the data were carried and the data were screened each column to grasp the meaning of each value in each column. This is a very crucial step to ensure the analysis will be conducted utilizes the right variables. Next, data pre-processing which divided into three steps such as data cleaning, data transformation and data preparation. Apart from that, in data preparation, the data were divided into train and test datasets with the ratio of 70:30 by using the `createDataPartition` command in R with p = 0.7. Ultimately, the train data (i.e. 15,090 observations) is utilized to build the model while the test data (i.e. 6,466 observations) is to validate the performance of the model when feed with a new dataset.

2.2 Exploratory Data Analysis

Exploratory data analysis is a method to analyze the characteristics of the dataset through visualization or descriptive statistics. This method helps in better understanding of the dataset and prepare the data for advanced analysis possibly. The distribution of the data can be represented in an appropriate visualization such as box plot or histogram for numeric variables and bar chart and pie chart for categorical variables.

2.3 Imbalanced Classification Treatment

Classification imbalanced dataset would produce biased model towards the majority class. Since, the train data for customers with no payment difficulties and have payment difficulties with classification imbalanced existed, thus, the down-sampling method was employed. The down-sampling technique works with randomly selected number of observations in the majority class to be equal to the number of observations in the minority class. Furthermore, down-sampling is appropriate to be implemented with a large dataset. This method not only solved the imbalanced issue rises but also helped in improving the runtime and storage issues during the model building process. In R, downSample that enables user to easily down-sampled the data in one linear command. Finally, the train dataset had an equal distribution of 1,727 in each class. The same configurations were applied to the test dataset to ensure the number of test dataset did not exceed the number of the training dataset (i.e. 739 observations in each class).

2.4 Model Building

The models employed for main analysis were random forest and improved machine learning approach i.e, XGBoost. Random forest is an ensemble of decision trees where the algorithm creates multiple decision trees and creates a forest that is more stable and provides better prediction. Extreme Gradient Boosting (XGBoost) is an improved algorithm based on the gradient boosting decision [18]. The advantage of XGBoost is that it is favorable amongst researchers, which its high execution speed and ability to outperform other methods.

$$Obj(\theta) = L(\theta) + \Omega(\theta) \tag{1}$$

Equation 1 is the objective function of XGBoost where L indicates the loss function which it was used to measure the performance of the training data. While Ω represents the regularization term which measures the model's complexity and aims to avoid overfitting. The model's complexity can be measured using the Equation 2 as below, where T represents the number of leaves of the tree while ω represents the weight of each leaf.

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{i=1}^{T} \omega_i^2 \tag{2}$$

2.5 Model Comparison and Assessment

To compare the performance of the prediction models for random forest and XGBoost, confusion matrix was used to compute the assessment metrics such as accuracy, precision, recall and F1-score. For accuracy, it can be computed by using (TP+TN)/(TP+TN+FP+FN), it helped to show the ability of the model to predict classes correctly. In addition, Receiver Operating Characteristics (ROC) curve also be utilized as a visualization approach to evaluate the model's performance. Area Under the Curve (AUC) was measured to compare the performance between the models, where model with higher AUC is better.

2.6 Feature Important Analysis

To determine the influential contributing factors in the default of loan repayment, dataset in XGBoost was chosen due to the best predictive model. In XGBoost library, this argument is useful to list out all the important features in the model, which `is xgb.importance()`.

3 RESULTS AND DISCUSSION

3.1 Characteristics on the Loan Default Dataset

Exploratory data analysis was conducted on the train dataset to further understand the behavior and distribution of each of the variables. The dataset consisted of continuous variables and categorical variables, then it can be represented in different ways such as histogram and bar chart respectively.

As a result, age of the customers between 30 to 50 years old was slightly skewed where most of them are at the time of their application. Most of the customers have been with their company for less than 10 years in the length of service of the customers' current employment as at the time of application and it resulted to a right-skewed histogram. The majority of the customers were female with 63.99%. However, the number of male customers having payment difficulties was higher with 14.17% against 9.91% of female customers. Most of the customers are married, whilst 2,302 of the customers are single. Other customers are divorced. Out of three classes, most of the customers with payment difficulties are single. It is because single people are younger with no stable income, hence led to higher default rate. Only 12 customers who have highest education level which is degree level whilst others are lower than that. Due to that, the customers with lower secondary qualification were reported the highest default rate amongst all the classes and it may cause them by having difficulties to meet their payment obligation. Due to limited space, the authors cannot afford to include data visualization in this article.

3.2 Imbalanced Classification Treatment

Imbalanced classification frequently leads to building a bias prediction model, where it produces poor performance on the minority class. Thus, it is vital to resolve this issue prior to model building. Before down-sampling was applied, shown in Figure 2, it was found that imbalanced classification existed in the dataset whereby only 11% which were 1,727 out of 15,090 observations were having payment difficulties whilst the remaining of 13,363 (88.6%) not have payment difficulties. Aforementioned, after down-sampling was applied to the train dataset, the TARGET variable was equally distributed with 1,727 observations in each class.

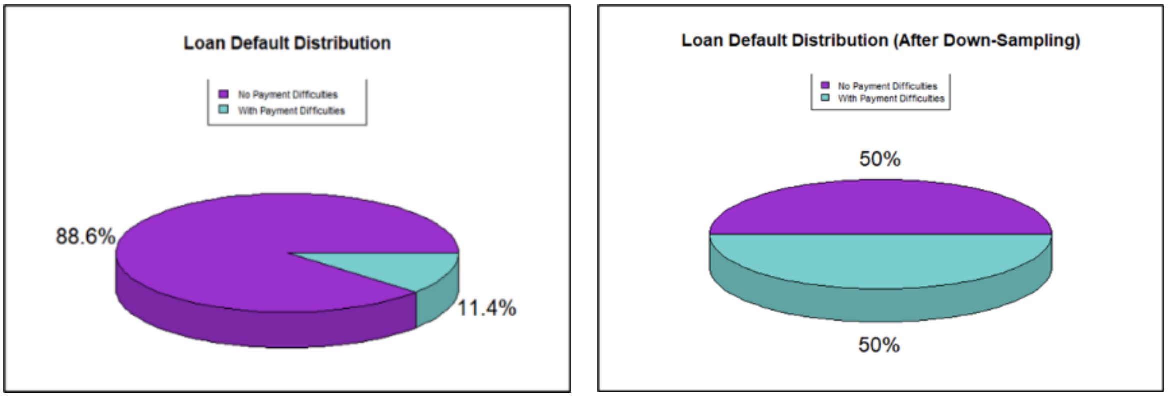

Figure 1 : Pie Chart of Loan Default Distribution (before and after down-sampling)

3.3 Random Forest for Model Building

Random forest model was built using the train dataset with three `mtry` and 1,500 `ntree`. Confusion matrix was used to predict whether the customers have payment difficulties or not. Table 1 shows the result of confusion matrix for random forest model by using test dataset.

Table 1 : Confusion matrix for Random Forest model and XGBoost Model

		Random Forest		**XGBoost**	
		Predicted			
		1	0	1	0
Actual	1	459	280	487	252
	0	308	431	313	426

As a result, 459 observations of the positive class were accurately predicted where both actual and predicted outcomes of customers having payment difficulties. Besides, 431 observations of the negative class were accurately predicted by the model of having both actual and predicted outcome of customers did not have payment difficulties. However, the model incorrectly predicted 280 observations of the positive class as it has classified those customers into outcome did not have payment difficulties, whilst 308 observations of the negative class was falsely classified into customers have payment difficulties. Comprehensively, the random forest model was able to predict 60.22% of the observations correctly. As mentioned by [19], the accuracy rate obtained was not far off from the accuracy that was documented in the study, where they utilized the same dataset to build random forest model.

3.4 XGBoost for Model Building

Extreme Gradient Boosting (XGBoost) model is different from random forest since it has more parameters that can be tuned by the users. Table 1 is the confusion matrix for XGBoost model by using test dataset. There are 487 observations of the positive class were accurately predicted for the customers have payment difficulties. However, the model was able to predict accurately of 426 observations of customers did not have payment difficulties, for instances, the negative class was lower as compared to random forest model. In other hand, there are 252 observations were inaccurately predicted by the model in the positive class was wrongly predicted as good customers. In contrary, 313 of the customers who did not have payment difficulties were falsely classified as bad customers by the model, which was higher than that predicted by random forest model. According to [7], false positive is more favourable as compared to false negative as misclassification of bad customers would incur more loss to the financial institutions. Inclusively, the XGBoost model was able to achieve higher accuracy rate of 61.77% as compared to random forest model.

3.5 Model Comparison and Assessment Results

Table 2 : Model Comparison and Assessment for Random Forest Model and XGBoost Model

	Random Forest	XGBoost
Accuracy	60.22%	61.77%
Precision	59.84%	60.88%
Recall	62.11%	65.90%
FI-Score	60.96%	63.29%

Table 2 shows the result of the model comparison between random forest and XGBoost. Generally, it shows that XGBoost was able to produce a predictive model with better performance in predicting the defaulted loans than random forest model. It is because XGBoost is able to produce higher accuracy rate of 61.77% compared to random forest model with 60.22%. While, for precision it can be seen that XGBoost was produced slightly higher for the positive class with 60.88% as compared to random forest. Next, XGBoost is outperformed compared to random forest as the recall rate and F1-Score rate is higher with 65.90% and 63.29%, respectively. Besides those rate, Receiver Operating Characteristics (ROC) curve of the two models as well as Area Under the Curve (AUC) value are executed. Figure 2 and Table 3 are the results of these method to examine the model performance.

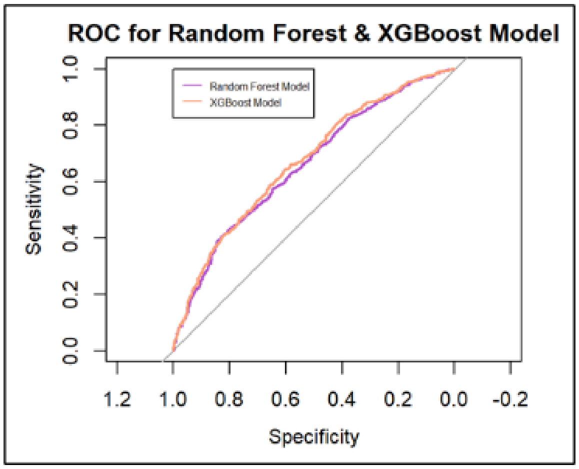

Figure 2 : ROC Curve of Random Forest Model and XGBoost Model

Table 3 : Area Under Curve of Random Forest Model and XGBoost Model

	Random Forest	XGBoost
AUC	0.6581	0.6709

Based on ROC, it resulted that the ROC of the random forest model was closer to the baseline as compared to the ROC of the XGBoost model. Whilst, the AUC value for XGBoost model is high by 1.28% from 0.6709 compared to the AUC for random forest model. Based on all criteria taken, it can be decided that XGBoost has better performance in predicting the model or the case study.

3.6 Feature Importance Analysis

Since, the XGBoost is the outperformed as compared to random forest model. All the variables in XGBoost were used in determining the influential contributing factors in the default of loan repayment. Based on result in Figure 3, there are 10 most importance features to predict the loan repayment in the XGBoost model. The most important feature is length of service with the Gain value of more than 0.25, followed by credit amount of the loan and also name education type, also the information on the customers' previous loan. It worth to highlight that the customers' demographic profile such as the duration of customers current employment and their highest education were ranked amongst the most important features in the model. Due to that, it is proven that the customers' demographic information was useful in determining the customers' ability to repay their loan.

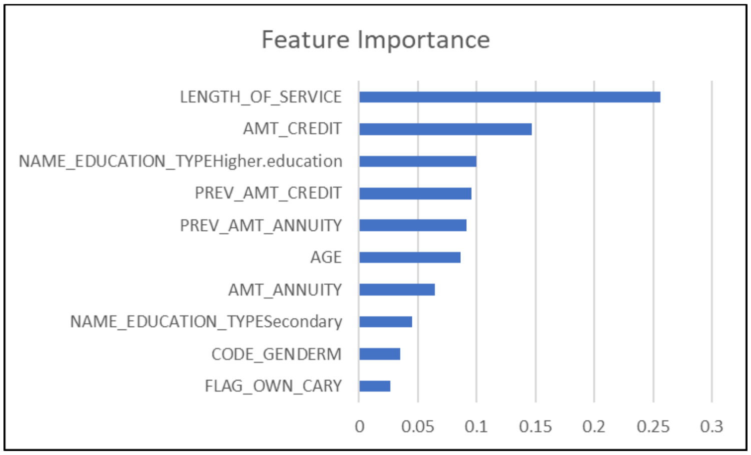

Figure 3 : Bar Chart of Feature Importance

4 CONCLUSION

The researchers and financial institutions have explored a lot of machine learning algorithms that enable to produce predictive models with improved performance, especially in risk management. In addition, this study used random forest and XGBoost to determine the best model in predicting the defaulted loans. Down-sampling used in this study helped to reduce the number of observations in the majority class so that it equates the number of observations in the minority class by randomly choosing the observations and avoiding from being biased. Thus, XGBoost is identified as a better performance in predicting the defaulted loans rather than random forest model. It is because, it was found to have good accuracy, precision, recall, F1-score and ROC as well as AUC compared to random forest. In addition, the gain matrix shows that demographic profile such as the customers' length of service with their current employer, level of education as well as their age was influential in determining their repayment ability. The findings of the study indicated that the predictive models, i.e. random forest and XGBoost, it was found that XGBoost was able to perform slightly better than random forest based on the result of all of the assessment matrices. It is found that the highest accuracy obtained by the best predicting model was 61.77%, with the application of random forest as well as XGBoost in predicting the loan default. Moreover, the comparison was done with some previous studies conducted by [19] and [20] where they have utilized both logistic regression and random forest, using the same loan default dataset. Both of their studies have recorded the same result of which logistic outperformed random forest in term of the accuracy where [19] recorded 69.34% accuracy of logistic against 70.60% recorded by [20]. For random forest model, [20] recorded higher accuracy of 68.40%, whilst [19] only managed to obtain a predictive model with 63.51% accuracy. The result obtained was slightly lower as compared to some findings reported by previous researchers which could be due to the application of lesser number of observations utilized in this study. Moreover, the programming package and environment constraints are also contributed to limitation in handling huge dataset. Nonetheless, the result obtained was still comparable and not very far off than that reported by previous studies.

It is recommended to the future studies, to explore and assess other models and algorithms using the same loan default dataset. As a result, the improvement on the existing models will be identified by considering different parameter tuning as well as feature selection. Moreover, it is recommended to explore other programming softwares that able to handle larger dataset such as SAS and python to enable more training dataset can be used in training the model. The better model performance can be achieved by this way as machine learning algorithm relies on the train dataset.

ACKNOWLEDGEMENT

The authors wish to thank to Universiti Teknologi Mara (UiTM) for giving the opportunity to complete part of this research project under Master of Science in Applied Statistics program. They would also like to express our appreciation to the ISTEC-CoSQA 2023 for the opportunity to present our research findings. Other than that, they also acknowledge their sincere appreciation to the reviewers for their valuable suggestion and remarks to improve the manuscript.

REFERENCES

[1] J.M. Zainol, A.M. Nor, S.N. Ibrahim, and S. Daud. "Macroeconomics Determinants of Non-Performing Loans in Malaysia: An ARDL Approach," *International Journal of Academic Research in Business and Social Sciences*, vol. 8, no.10, 2018. doi:10.6007/ijarbss/v8-i10/4773.

[2] Z. Khairudin, N.A. Abdul Razak, H. A., Abd Rahman, N., Kamarudin, and N.A., Abd Aziz.

"Prediction of diabetic retinopathy among type ii diabetic patients using data mining techniques,". *Malaysian Journal of Computing*, vol. 5, no. 2, pp. 572-586, 2020.

[3] M.N. Abdullah, Y.B. Wah, A.B. Majeed, Y. Zakaria, and N. Shaadan. "Identification of Blood-Based Multi-Omics Biomarkers for Alzheimer's Disease Using Firth's Logistic Regression," *Pertanika Journal Science & Technology,* vol. 30, no. 2, pp. 1197-1218, 2022.

[4] N.A.M. Salim, Y.B. Wah, C. Reeves et al.. "Prediction of dengue outbreak in Selangor Malaysia using machine learning techniques," *Scientific Reports,* no. 11, pp. 939, 2021.

[5] W.N. Shaziayani, A.Z. Ul-Saufie, S. Mutalib, N. Mohamad Noor, and N.S. Zainordin. "Classification Prediction PM10 Concentration Uisng a Tree-Based Machine Learning Approach," *Atmosphere,* no. 13, pp. 538, 2022. http://doi.org/10.3390/ atmos13040538.

[6] S.Z. Ramlan, and S. Mohd Deni. "Rainfall Prediction in Flood Prone Area Using Deep Learning Approach. In: A. Mohamed, Y.B. Wah, J.M. Zain, M.W. Berry (eds) Soft Computing in Data Science. SCDS 2021," *Communications in Computer and Information Science,* vol. 1489, Springer, Singapore, 2021. https://doi.org/10.1007/978-981-16-7334-4_6.

[7] Leon. "Predictive Modelling for Loan Defaults," *UCLA Electronic Theses and Dissertations,* 2019. https://escholarship.org/uc/item/3rs9b3d6.

[8] Y. Chen, W. Zheng, W. Li, and Y. Huang. "Large group activity security risk assessment and risk early warning based on random forest algorithm," *Pattern Recognition Letters,* no. 144, pp. 1-5, 2021. https://doi.org/10.1016/j.patrec.2021.01.008..

[9] M. Madaan, A. Kumar, C. Keshri, R. Jain, and P. Nagrath. "Loan default prediction using decision trees and random forest: A comparative study," *IOP Conference Series: Materials Science and Engineering,* vol. 1022, no. 1, 2021. https://doi.org/10.1088/1757-899X/1022/1/012042.

[10] Q. Truong, M. Nguyen, H. Dang, and B. Mei. "Housing Price Prediction via Improved Machine Learning Techniques," *Procedia Computer Science,* no. 174, pp. 433-442, 2020. https://doi.org/10.1016/j.ins.2021.01.059.

[11] Basel Committee on Banking Supervision. "Guidelines for definitions of non-performing exposures and forbearance." Retrieved from https://www.bis.org bcbs/publ/d367.pdf.

[12] D.P. Louzis, A.T. Vouldisn, and V. L. Metaxas. "Macroeconomic and bank-specific determinants of non-performing loans in Greece: A comparative study of mortgage, business and consumer loan portfolios," *Journal of Banking & Finance,* vol. 36, no. 4, pp. 1012-1027, 2012.

[13] J. Wan, Z.L. Yue, D.H. Yang, Z. Yu, L. Jiao, L. Zhi, and J. Liu. "Prediction non performing loan of business bank with data mining techniques," *International Journal of Database Theory and Application,* vol. 9, no. 12, pp. 23-34, 2016.

[14] J. Bhargava, and P.R. Masuku. "Identifying the factors responsible for loan defaults and classification of customers using SAS® Enterprise Miner." https://www.sas.com/content/dam/SAS/en_us/doc/event/analytics-experience-

2016/identifying-factors-responsible-loan-defaults-classification-customers-using-sas-em.pdf.

[15] L. Zhu, D. Qiu, D. Ergu, C. Ying, and K. Liu. "A study on predicting loan default based on the random forest algorithm," *Procedia Computer Science,* no. 162, pp. 503-513, 2019. https://doi.org/10.1016/j.procs.2019.12.017.

[16] R. Odegua. "Predicting Bank Loan Default with Extreme Gradient Boosting." https://doi.org/10.48550/arXiv.2002.02011.

[17] X. Ma, J. Sha, D. Wang, Y. Yu, Q. Yang, and X. Niu. "Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning," *Electronic Commerce Research and Applications,* no. 31, pp. 24-39, 2018. https://doi.org/10.1016/j.elerap.2018.08.002.

[18] C. Yang, M. Chen, and Q. Yuan. "The application of XGBoost and SHAP to examining the factors in freight truck-related crashes: An exploratory analysis," *Accident Analysis and Prevention,* no. 158, 2021. https://doi.org/10.1016/j.aap.2021.106153.

[19] Y. Liang, X. Jin, and Z. Wang. "Loanliness: Predicting Loan Repayment Ability by Using Machine Learning Methods." https://github.com/Yiyun-Liang/loanliness.

[20] Z. Qiu, Y. Li, P. Ni, and G. Li. "Credit risk scoring analysis based on machine learning models," *Proceedings – 2019 6th International Conference on Information Science and Control Engineering, ICISCE 2019,* pp. 220-224, 2019. https://doi.org/10.1109/ICISCE48695.2019.00052.