Journal of Engineering and Technology

ISSN 2180-3811 eISSN 2289-814X https://jet.utem.edu.my/jet/index

PREDICTING THE LOAN DEFAULT USING MACHINE LEARNING ALGORITHMS: A CASE STUDY IN INDIA

W. T. Loo[1], K. W. Khaw*[1], X. Y. Chew[2], A. Alnoor[3], S.T. Lim[1]

[1] School of Management, Universiti Sains Malaysia, 11800 USM Penang, Malaysia

[2] School of Computer Sciences, Universiti Sains Malaysia, 11800 USM Penang, Malaysia

[3] Management Technical College, Southern Technical University, Basrah, Iraq

corresponding: khaiwah@usm.my

Article history:
Received Date:
4 February 2023
Revised Date:
24 August 2023
Accepted Date:
22 September 2023

Keywords: Bank, India, Machine Learning Algorithms,

Abstract— The main income of banks was generated from mobilizing the deposits to borrow to applicants. Although applying for loans is becoming common, banks still need to take the risk that the applicants may have a loan default. In this study, the objectives are to predict the risk of loan default using 6 types of machine learning (Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbor, Support Vector Machine, and Naïve Bates), compare the machine learning algorithms to choose the most suitable algorithms for predicting the

Prediction, Risk of Loan Default	risk of loan default, and help the decision maker in approving or rejecting the loan requests. The dataset is focused on India with their behaviour to determine their risk. Using the Jupyter Notebook (Python) to build the model and evaluate each model. There are 5 types of evaluation metrics (Accuracy, Precision, Recall, F1-Score and Average) are used to determine the champion model among the six machine learning algorithms. In this study, K-Nearest Neighbor is the champion model because this model scored the highest in all the evaluation metrics, which is 0.89. Although machine learning algorithms can help to determine the risk of flagging, the decision maker should take some actions to decrease the risk of loan default such as creating a clear plan for the payment reminders and providing a convenient way.

I. Introduction

The main business of a bank is depositing the mobilization, which the bank lends to borrowers, which is the main income-generating [1]. Besides, rapid credit growth risks banks in the following year [2]. The bank, one of the financial institutions, controls the country’s economic development, such as agriculture and industry [3]. Hence, bad borrowing may make banks risky and affect the economy.

According to a survey of banking professionals, 72% of participants believe India's NPA crisis will worsen [4] [5]. NPA refers to Mounting non-performing assets, attracting the attention of different sectors, such as economists, stakeholders, and decision-makers. This shows that the economic environment in

India should be monitored compactly. Two types of information are needed to predict applicants' loan default risk [6]. They are financial information such as their income or the relevant financial status, and information such as their age and the city.

Loan default happens when one misses payments for a specified period of time. There are various loans, such as student loans, mortgages, credit card loans, auto loans, secured personal loans, and unsecured personal or business loans [7]. In other words, loan default risk is defined as the risk of loss due to the applicants' failure to deliver the contract on time [8]. In the loan default prediction process, the institute may suffer significant financial losses when a defaulter is mistakenly categorized as a non-defaulter during the default prediction process. On the contrary, when a non-defaulter is identified as a defaulter, it will disrupt the pool of high-quality clients [9].

Although applying for loans has become common, banks still need to exercise caution when granting loans to the public to protect themselves from loan defaults. If an applicant defaults on the loan, the bank must take action, which consumes time, energy, and money. Hence, they need to establish a risk assessment system that can assist them in evaluating the ability of loan applicants to repay. However, processing a large volume of data manually is impractical. Implementing machine learning methods is the most suitable approach for them to build a model and predict the risk associated with loan applicants.

II. Literature Review

A. Related Works

The relation between the applicants and the risk of default is high, which means the risk can be predicted based on their characteristics to estimate the distance [10].

Random Forest (RF), Extreme Gradient Boosting Tree (XGBT), Neural Network (NN), and Gradient Boosting Model (GBM) are compared [11]. Through their paper, the result shows the borrowers who are passed in asset and the individual credit will be less risky to have loan default. The superior machine learning algorithm is Random Forest which gains 0.984 in accuracy.

In the same dataset, Random Forest gains higher accuracy than Decision Tree, which was proved [12]. In this paper, the authors show that owned assets will be less likely to have loan default.

Journal compared with Logistic Regression, Random Forest, Decision Tree, and Support Vector Machine. Random forest is the superior machine learning algorithm which gains a 95% accuracy rate compared with other fields [13].

The borrowers who gain higher income will be less likely to have loan default, and their result shows that Random Forest is the champion machine learning algorithm compared to other fields [14].

B. Machine Learning Algorithms

Logistic regression can construct a separating hyperplane between two datasets [15]. It expresses the distance from the hyperplane as a probability of class membership. When the variables are fewer, the model of logistic regression is considered low complexity. Its process is to fit the data into a logistic curve and the event occurring probability can be evaluated using the relationship between the independent factors and categorical dependent variables [16].

The Decision Tree classifier can be expressed as a recursive partition of the instance space [17]. Each node in the decision tree indicates a feature of the instance that would be classed, and each branch provides a potential value for the node [18]. Its prediction accuracy is normally lower than the other machine learning algorithms. It also costs a long training time and high computation [19].

Random forest is the machine learning algorithm that generates many decision trees based on the random subsamples of the training set. The features in the trees will be randomly varied [20]. Good binary splits are the two daughter nodes receiving data from the parent tree node . This ensures that the homogeneity in the daughter nodes and the parent node can be improved when pushing the data to the daughter nodes. Due to the selection being random, the different random forest is different in randomness [1].

Support Vector Machines (SVM) gain the best classification function of training data among all the machine learning algorithms [21]. Then, training the model is easy, and it scales relatively well to the high dimensional data [22] [23]. The trade-off between the model complexity and the error can be easily controlled. It can proceed regardless of the continuous or categorical data. The prediction accuracy is very high and has good generalization capability with limited training samples [22] [24].

The K-Nearest Neighbors (KNN) algorithm is one of the simplest classification algorithms to solve classification and regression problems [25] [26]. “K” is a number that refers to the amount of nearest neighbors used, which can be calculated using the given value's upper limit or directly defined in the object builder [27]. The choice boundary is implicitly computed by its function, and the decision boundary can also be computed [21].

Naïve Bayes is a classification method that applies Bayes' rule under the presumption of predictor independence [28]. The benefit of Nave Bayes is that it does calculations quickly and with a wide range of capabilities [21][29]. If conditional Independence is true, Naïve Bayes could produce excellent results [30]. It can also manage missing values for attributes [30]. When running on a big scale, Nave Bayes can occasionally outperform other algorithms [18].

III. Methodology

The process of this work is shown in Figure 1.

A. Data Retrieval and Data Pre-processing

The dataset implemented in this work was retrieved from a public website, named Kaggle. The dataset belonged to a Hackathon organized by “Univ.AI”.

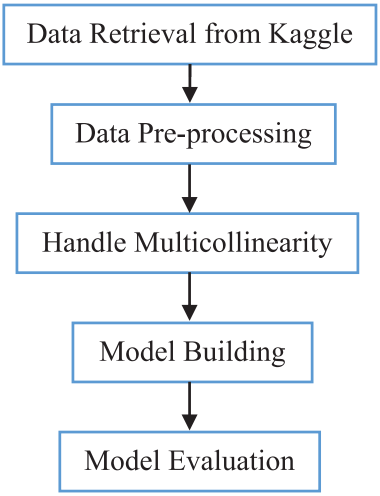

Figure 1: Framework

Before building the models, there are some pre-processing is run. First, handling the missing value. Make sure there is no missing value in each column. Deleting unnecessary variables such as ID. Checking the outliers. Feature scaling and standardization and encoding categorical variables to change the categorical variables into numerical variables.

B. Preview of Multicollinearity

Multicollinearity occurs when there are two or more independent variables are correlated [15]. In this study, the variables have low correlations with the risk flag. Multicollinearity confuses tests of significance in this way. The more characters that are considered and the stronger the correlations between them, the less probable it is that the null hypothesis [22].

C. Model Building

The dataset was split into two as the training dataset and the testing dataset. Then, use the libraries of each model to build the models. Lastly, evaluation of the models.

D. Model Evaluation

The performance of each model was evaluated in terms of several metrics, such as accuracy, precision, recall, F1-score, and average score.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1-Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

$$Average\ Score = \frac{Sum\ of\ Scores}{4} \quad (5)$$

where:

TP = True Positive, the defaulter is truly a defaulter.

TN = True Negative, the non-defaulter is truly a non-defaulter.

FP= False Positive, the non-defaulter is wrongly identified as a defaulter.

FN= False Negative, the defaulter is wrongly identified as a non-defaulter.

IV. Results and Discussion

A. Result

The machine learning model that gained the highest accuracy is KNN, which is 0.89 (rounded up as 89%). The rest of the models gained 0.88 (rounded up as 88%).

Decision Tree and KNN have the same precision value, which is 0.89 becoming the highest among the six models, while Random Forest gained the second highest, which is 0.86. The rest of the models are the same 0.77 becomes the lowest.

KNN is still the highest among the six models in the recall part. KNN achieved 0.89 (rounded up as 89%), while the rest of the models are 0.88 (rounded up as 88%).

KNN is the highest in F1-score, which is 89%, while the second highest is Decision Tree which is 88%. The model which gained the lowest score is Logistic Regression, Support Vector Machine and Naïve Bayes, which only gained 82%.

B. Discussion

The result displays that almost all models perform well in the loan default prediction. However, KNN slightly outperformed the others. This may be due to its non-parametric characteristics and its outliers' robustness [31].

Moreover, KNN is also well-known for its simplicity in model building. Therefore, KNN not only possesses the highest value in all the evaluation metrics but also is the most effortless model building in the loan default prediction.

C. Champion Model

The champion model is KNN since the metrics scores of every metric are the highest (89%) while the rest of the models gained lower scores.

Table 1: Performance Result of Models

Models	Evaluation Metrics				
	Accuracy	Precision	Recall	F1-Score	Average
LR	0.88	0.77	0.88	0.82	0.84
DT	0.88	0.89	0.88	0.88	0.88
RF	0.88	0.86	0.88	0.83	0.86
SVM	0.88	0.77	0.88	0.82	0.84
KNN	0.89	0.89	0.89	0.89	0.89
NB	0.88	0.77	0.88	0.82	0.84

V. Implication

A. Provide a Convenient Way

Nowadays, in a cashless era, e-wallets and e-banking are common and popular. As we know, most senior citizens are not proficient in using digital banking to transfer money. Hence, the counter services should be run normally. However, some of the senior citizens have the health problems such as cannot walk normally. The bank should ask the applicant to collab the name with someone like their children so that the children are responsible for paying back the money for the applicants. E-wallet for loan payment has not been carried out yet, but this may be a choice for the appliers to repay the loan.

VI. Conclusion

The analysis of loan default prediction models offers important new perspectives on the crucial issue of determining applicants' probability of defaulting on their loans. The banking sector is crucial to a country's economic growth since it directs deposits into leading operations. Rapid loan expansion, however, is dangerous for the economy as a whole as well as banks. The rise of non-performing assets (NPAs) has drawn attention from various stakeholders. This underscores the importance of monitoring the economic climate, especially in nations like India, where NPAs are a growing concern. Loan default happens when someone does not make their contractually required payments on time. Given the impracticality of manually processing massive volumes of data, the machine learning approach is the most practical way to create models for predicting loan default risk. K-Nearest Neighbors (KNN) exhibited the highest accuracy at 89%, slightly outperforming the other models. KNN's non-parametric nature and robustness to outliers likely contributed to its superior performance. In addition, KNN's ease of model construction adds to its appeal, making it the most efficient model as well as the easiest to use in the context of loan default prediction. KNN is, therefore, without a doubt, the best model for predicting loan default. Nevertheless, in future work, it is recommended to include feature selection in the data pre-processing process to prevent over-fitting. Meanwhile, ensemble models or hybrid models can also be applied to undergo the prediction.

VII. Acknowledgement

This work is supported by the Ministry of Higher Education Malaysia, Fundamental and Research Grant Scheme under Grant No. FRGS/1/2019/STG06/ USM/ 02/6 for the project entitled “A New Hybrid Model for Monitoring the Multivariate Coefficient of Variation in Healthcare Surveillance”.

VIII. References

[1] G. Biau, L. Devroye and G. Lugosi, "Consistency of Random Forests and Other Averaging Classifiers," *Journal of Machine Learning Research 9,* pp. 2015-2033, 2008.

[2] D. Foos, L. Norden and M. Weber, "Loan growth and riskiness of banks," *Journal of Banking and Finance,* vol. 34, no. 12, pp. 2929-2940, 2010.

[3] M. J. Patwary, M. S. B. Alam, S. Akter and A. R. Karim, "Bank Deposit Prediction Using Ensemble Learning," *Artificial Intelligence Evolution,* pp. 42-51, 2021.

[4] S. Bhadury and B. Pratap, "India's Bad Loan Conundrum: Recurrent Concern for Banking System Stability and the Way Forward," *International Symposia in Economic Theory and Econometrics,* vol. 25, pp. 123-161, 2018.

[5] M. L. Bhasin, "Unmasking Rising NPAs: Can the Indian Banking Sector Overcome this Phase?," *International Journal of Management and Social Science Research,* vol. 6, no. 3, pp. 5-19, 2017.

[6] S. Angilella and S. Mazzu, "The financing of innovative SMEs: A multicriteria credit rating mode," *European Journal of Operational Research,* vol. 244, no. 2, pp. 540-554, 2015.

[7] E. F. Malik, K. W. Khaw, B. Belaton, W. P. Wong and X. Y. Chew, "Credit card fraud detection using a new hybrid machine learning architecture," *Mathematics*, vol. 10, 1480, 2022.

[8] K. Yuan, G. Chi, Y. Zhou and H. Yin, "A novel two-stage hybrid default prediction model with k-means clustering and support vector domain description," *Research in International Business and Finance,* vol. 59, p. 101536, 2022.

[9] M. Papouskova and P. Hajek, "Two-stage consumer credit risk modelling using heterogeneous ensemble learning," *Decision Support Systems,* vol. 118, pp. 33-45, 2019.

[10] J. W. Lee and S. Y. Sohn, "Evaluating borrowers' default

risk with a spatial probit model reflecting the distance in their relational network," *PLoS ONE,* vol. 16, no. 12, p. e0261737, 2021.

[11] J. Xu, Z. Lu and Y. Xie, "Loan default prediciton of Chinese P2P market: a machine learning methodology," *Scientific Report,* vol. 11, no. 1, 2021.

[12] M. Madaan, A. Kumar, C. Keshri, R. Jain and P. Nagrath, "Loan default prediction using decision trees and random forest: A comparative study," in *Materials Science and Engineering*, 2021.

[13] L. Zhu, D. Qiu, D. Ergu, C. Ying and K. Liu, "A study on predicting loan default based on the random forest algorithms," in *Procedia Computer Science*, 2019.

[14] J. I. Daoud, "Multicollinearity and Regression Analysis," in *Journal of Physics: Conference Series*, 2018.

[15] S. Dreriseitl, L. Ohno-Machado, H. Kittler, S. Vinterbo, H. Billhardt and M. Binder, "A comparison of machine learning methods for the diagnosis of pigmented skin lesions," *Journal of Biomedical Informatics,* vol. 34, no. 1, pp. 28-36, 2001.

[16] H. A. Park, "An introduction to logistic regression: From basic concepts to interpretation with particular attention to nursing domain," *Journal of Korean Academy of Nursing ,* vol. 43, no. 2, pp. 154-164, 2013.

[17] V. Nasteski, "An overview of the supervised machine learning methods," *Horizons.B.,* vol. 4, pp. 51-62, 2017.

[18] F. Osisanwo, J. Akinsola, O. Awodele, J. Hinmikaiye, O. Olakanmi and J. Akinjobi, "Supervised Machine Learning Algorithms: Classification and Comparison," *International Journal of Computer Trends and Technology,* vol. 48, no. 3, pp. 128-138, 2017.

[19] S. D. Jadhav and H. Channe, "Comaprative Study of K-NN, Naive Bayes, and Decision Tree Classification Techniques," International Journal of Science and Research, 2013.

[20] S. Dreiseitl, L. Ohno-Machado, H. Kittler, S. Vinterbo, H. Billhardt and M. Binder, "A Comparison of Machine Learning Methods for the Diagnosis of Pigmented Skin Lesions," *Journal of Biomedical Informatics,* vol. 34, pp. 28-36, 2001.

[21] P. Singh, S. Verma, I. Khan and S. Sharma, "Machine Learning: A Comprehensive Survey on Existing Algorithms," *Journal of Computer Science Engineering and Software Testing,* vol. 7, no. 3, pp. 1-9, 2021.

[22] A. E. Mohamed, "Comparative Study of Four Supervised Machine Learning Techniques for Classification," *International Journal of Applied Science and Technology,* vol. 7, no. 2, pp. 5-18, 2017.

[23] T. Wuest, D. Weimer, C. Irgens and K.-D. Thoben, "Machine learning in manufacturing: advantages, challenges, and applications," *Production & Manufacturing Research,* vol. 4, no. 1, pp. 23-45, 2016.

[24] Y. Shao and R. S. Lunetta, "Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited training data points," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 70, pp. 78-87, 2012.

[25] B. Mahesh, "Machine Learning Algorithms - A Review," *International Journal of Science and Research,* 2018.

[26] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," *Computers in Biology and Medicine ,* vol. 136, p. 104672, 2021.

[27] X. Luo, H. Li, H. Wang, Z. Wu, F. Dai and D. Cao, "Vision-based detection and visualization of dynamic workspaces," *Automation in Construction,* vol. 104, pp. 1-13, 2019.

[28] I. H. Sarker, M. M. Hoque, M. K. Uddin and T. Alsanoosy, "Mobile Data Science and Intelligent Apps: Concepts, AI-Based Modeling and Research Directions," *Mobile Networks and Applications,* vol. 26, pp. 285-303, 2021.

[29] A. C. Y. Hong, K. W. Khaw, X. Y. Chew and W. C. Yeong, "Prediction of US airline passenger satisfaction using machine learning algorithms. *Data Analytics and Applied Mathematics*, vol. 4, pp. 8-24, 2023.

[30] D. M. Farid, L. Zhang, C. M. Rahman, M. Hossain and R. Strachan, "Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks," *Expert Systems with Applications,* vol. 41, pp. 1937-1946, 2014.

[31] T. Srivastava, "Analytics Vidhya - A Complete Guide to K-Nearest Neighbors," 14 7 2023. [Online]. Available: https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/.