Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 01, 2024, DOI 10.54209/infosains.v14i01
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

An implementation of machine learning on loan default prediction based on customer behavior

[1]Robi Aziz Zuama, [2]Nurul Ichsan, [3]Achmad Baroqah Pohan, [4]Mohammad Syamsul Azis, [5]Mareanus Lase
[1,2,3]Fakultas Teknik & Informatika, Universitas Bina Sarana Informatika, [4,5]Fakultas Teknologi Informasi, Universitas Nusa Mandiri

Article Info
Keywords:
Loan Default Prediction,
Machine Learning,
Customer Behavior Analysis,
Financial Institution Challenges,
Matrix Evaluation

ABSTRACT
In the banking sector, loans have become a key component that steers the economy, encourages company expansion, and directly impacts the growth of a nation's economy. Banks must evaluate borrowers' ability to repay loans given the inherent risks involved in order to reduce the likelihood of default. In particular, machine learning (ML) has shown promise as a revolutionary tool for loan default prediction using advanced methodologies to examine historical data relating to customer behavior, this study investigates the application of machine learning (ML) in forecasting loan outcomes. The results show that XGBoost performs better than other machine learning algorithms, with an accuracy rate of 89%. Random forest and logistic regression come in second and third, respectively, with 88% accuracy. KNN and decision trees come next, both with somewhat lower accuracy rates (87%). By incorporating consumer behavior domain variables, this study fills in the gaps in the literature and offers a more thorough understanding of loan projections. In order to improve model performance and strengthen the predictive power of machine learning algorithms in loan scenarios, further research incorporating trials to optimize algorithm parameters is necessary as financial institutions continue to experience difficulties.

Corresponding Author:
Robi Aziz Zuama
Fakultas Teknik & Informatika, Universitas Bina Sarana Informatika
robi.rbz@bsi.ac.id

INTRODUCTION

Over the years, loans as a banking product have played a crucial role in driving the economic system [1], This is because loans contribute directly to the country's economic growth by assisting in the development of businesses [2][3]. As a financial institution, banking is essential to maintaining financial system stability and preventing excessive risk-taking. Since loans are a risky banking product, banks are obligated to assess the borrower's ability to repay the loan on time to minimize such risks [4] [5]. The Financial Services Authority (OJK) in Indonesia reported that the national non-performing loan value reached IDR 1.73 trillion in June 2023. Non-performing loans are assessed based on customers' failure to make payments more than 90 days past the due date, known as the >90 days past due or the default rate [6]. Therefore, it is crucial for banking institutions to conduct risk evaluation and borrower portfolio assessment to minimize the risk of default [7].

Machine learning (ML) has opened doors to innovation in various fields, including finance [8]. In the financial sector, machine learning is utilized to predict loan defaults. Using ML for prediction helps financial companies reduce the likelihood of losses due to defaulted loans [9], The adoption of ML in loan prediction provides valuable information for financial institutions to make better decisions in assessing credit risks and deciding to approve or reject loans [10].

Previous research has extensively implemented machine learning in loan prediction, such as the study [11] which used ML models for credit scoring and default prediction using NLP, embedding, autoencoder, and GBM techniques. For imbalanced samples, probability calibration, and credit ratings were established using the DE genetic algorithm. Techniques like SHAP and LIME improved interpretation. The study [12] analyzed loan approval predictions by considering demographic data and approval types using machine learning techniques. This study compared the efficiency of decision tree and logistic regression models based on client financial data. The results showed that the decision tree outperformed conventional techniques in predicting loan approvals. The study [13] proposed a better approach using machine learning (KNN, decision tree, SVM, and logistic regression). Jaccard similarity coefficient, log loss, and F1 score were used to test accuracy. This helps banks save time and resources by reducing the borrower verification process.

From these studies, it can be concluded that ML has significant potential in improving loan predictions. However, there are still shortcomings and challenges to overcome, such as choosing the appropriate algorithm and considering customer behavior factors such as marital status, existing loans, car ownership, profession, residence location, length of employment, and others.

In this research, we will explore the implementation of ML in predicting loans based on customer behavior. Through the use of advanced ML techniques, we will analyze historical data related to customer behavior. Thus, this research aims to contribute significantly to improving the efficiency and accuracy of the loan eligibility determination process.

METHOD

Data Description

In this study, the dataset used is public data obtained from Kaggle [14]. The number of variables utilized is 12, and the description of these variables can be seen in the table. The table provides details for a subset of the data,
Table 1 explaining the 12 variables.

Table 1. Data Descriptions

No	Column	Description	Type
1	Income	Income of the user	int
2	age	Age of the user	int

No	Column	Description	Type
3	experience	Professional experience of the user in years	Int
4	profession	Profession	String
5	married	Whether married or single	String
6	house_ownership	Owned or rented or neither	String
7	car_ownership	Does the person own a car	String
8	current_job_years	Years of experience in the current job	Int
9	current_house_years	Number of years in the current residence	Int
10	city	City of residence	String
11	state	State of residence	String
12	risk_flag	Defaulted on a loan	string

This study employs a dataset with 252,000 entries, consisting of 221,004 instances of no loan default and 30,996 instances of loan default, Figure 1. The diagram explains the comparison between defaulters and non-defaulters

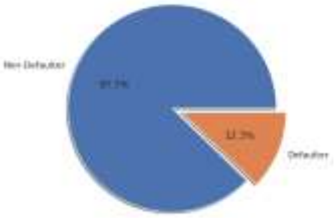


Figure 1. The diagram explains the comparison between defaulters and non-defaulters

Data Preprocessing

Data preprocessing is crucial as low-quality raw data can result in a decrease in prediction accuracy. Preprocessing involves handling missing or problematic data, selecting relevant features, and reducing duplicate values Additionally, a preprocessing task involves balancing classes. In this study, the oversampling method is employed, utilizing the SMOTE (Synthetic Minority Over-sampling Technique) method to balance loan default and non-default classes [15].

Machine Learning

1. Logistic Regression

Logistic regression is a statistical method used to model the relationship between one or more independent variables and a binary dependent variable [16]. Logistic regression is suitable for classification problems where the goal is to predict the probability of events in two categories [17]. Equation (1) Explains logistic regression

$$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}} \quad (1)$$

Here, P(Y=1) is the probability of a positive event, e is the natural logarithm base, β_0 to β_n are the model parameters, and X_1 to X_n are the independent variables

2. Decision Tree

Decision Tree is a machine learning algorithm that maps decisions based on a series of hierarchical rules formed from training data. Each node in the tree represents a decision based on a specific feature, leading to child nodes [18] [19]. The advantages of Decision Tree include ease of interpretation, ability to handle non-linear data, and not requiring specific distribution assumptions.

3. Random Forest

Random Forest is an ensemble machine learning algorithm that builds a large number of decision trees during training and combines their prediction results to enhance model accuracy and resilience. Random Forests work by combining decisions from many randomly created decision trees to provide more accurate predictions [20]. The advantages of Random Forest lie in high accuracy, resilience to overfitting, and excellent feature handling.

4. XGBoost

eXtreme Gradient Boosting (XGBoost) is a boosting algorithm that enhances the performance of machine learning models by combining predictions from a small number of decision trees. In XGBoost, the model is built gradually, and each tree focuses on overcoming the prediction errors of the previous model [21]. The objective function of XGBoost consists of two parts: loss and regularization. The general objective function is written as (1):

$$l(y,\hat{y}) + \Omega(f_k) \quad \text{.......................(2)}$$

l is the loss function (such as squared error for regression, log likelihood for classification), y_i is the model prediction, and $\Omega(f_k)$ is the regularization function for each tree.

5. Evaluation Metrics

In the context of using machine learning for classification, evaluation metrics are used to measure or evaluate the performance of a model in predicting the target class [22]. The most common matrices used, such as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$ Precision = $\frac{TP}{TP+FP}$

Recall or Sensitivity = $\frac{TP}{TP+FN}$ F1-Score = $2 \times \left(\frac{Presisi \times Recall}{Presisi+Recall}\right)$

Examining these metrics helps us understand how well our model can distinguish between positive and negative classes and how well they minimize prediction errors.

RESULTS AND DISCUSSION

In this study, the data is divided into two parts: training data and testing data. The training data comprises 70% of the total dataset, while the remaining 30% is used as testing data. To observe prediction outcomes, this research employs machine learning algorithms such as logistic regression, decision tree, random forest, and XGBoost. To evaluate the performance of the predictive models, the study applies an evaluation matrix to determine accuracy, precision, F1 score, and recall values.

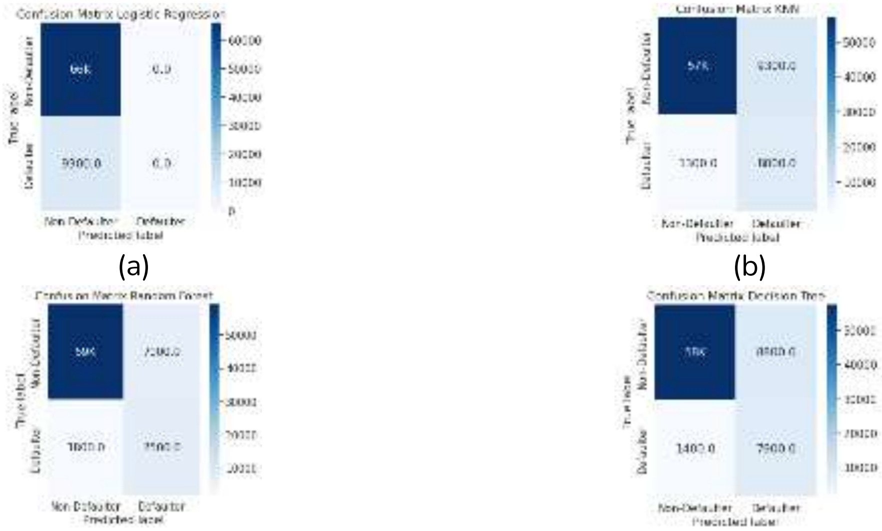
The performance of five machine learning algorithms is compared, revealing significant differences. Based on accuracy, the XGBoost algorithm demonstrates the highest performance compared to the other algorithms, achieving an accuracy of 89%. Logistic regression and random forest also exhibit satisfactory performance compared to

KNN and decision tree, with accuracy values of 88% for each algorithm. The accuracy of the decision tree is 87%, and finally, KNN achieves an accuracy of 86%. This study additionally discovers the performance of machine learning algorithms in terms of precision, recall, and F1 score. Table 2 presents the comparative performance results of the five machine learning algorithms, where all algorithms are evaluated for their ability to predict default or non-default instances.

Table 2. Comparison of classification report

Algorithms	Class	Precision	Recall	F1-score	Accuracy
Logistic Regression	Non-Defaulter	0.88	1.00	0.93	0.88
	Defaulter	0.00	0.00	0.00	
KNN	Non-Defaulter	0.98	0.86	0.92	0.86
	Defaulter	0.46	0.86	0.60	
Random Forest	Non-Defaulter	0.97	0.89	0.93	0.88
	Defaulter	0.51	0.81	0.63	
Decision Tree	Non-Defaulter	0.98	0.87	0.92	0.87
	Defaulter	0.47	0.85	0.61	
XGBoost	Non-Defaulter	0.97	0.91	0.94	0.89
	Defaulter	0.54	0.78	0.64	

Untuk dapat memvisualisakan hasil kinerja prediksi dari lima algoritma yang diusulkan, maka penggunaan confusion matrix bisa digunakan. Setiap nilai entri consudion matrix menunjukkan jumlah prediksi yang dibuat model ketika mengklasifikasikan kelas dengan benar atau salah. Gambar 2 (a) menunjukan confusion matrix logistic regression, gambar 2 (b) menunjukan confusion matrix KNN, gambar 2 (c) menunjukan confusion matrix random forest, gambar 2 (d) menunjukan decision tree dan gambar 2 (e) menunjukan confusion matrix XGBoost.

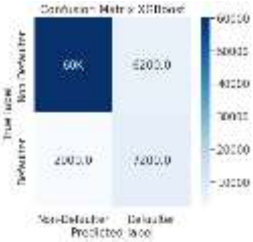


(a)

(b)

(c)

(d)



(e)

Figure 2. (a) Confusion Matrix Logistic Regression, (b) Confusion Matrix KNN, (c) Confusion Matrix Random Forest, (d) Confusion Matrix Decision Tree, (e) Confusion Matrix XGBoost

Sebagian besar penelitian dalam prediksi pinjaman menggunakan kecerdasan buatan hanya berfokus pada beberapa implementasi model dan menggunakan scoring, demografi calon nasabah dan jenis pinjaman. Namun, hanay sedikit penenelitian tentang prediksi pinjaman dengan variabel domain perilaku konsumen, seperti menyangkut pekerjaan, status keluarga dan lain-lain.

Penelitian ini mengimplemtasikan algortima machine learning untuk memprediksi pinjaman macet (gagal Bayar) atau lancar, dengan mempertimbangkan perilaku konsumen. Untuk mengevaluasi hasil kinerja prediksi algoritma mengguanakan confusion matrix, hasil penelitian menunjukan performa machine learning pada algoritma XGBoost menunjukan hasil yang baik, dengan tingkat akurasi 89%, sedangkan logistic regression dan random forest mempunyai tingkat akurasi yang sama yaitu 88%,, Decision tree sebesar 87% dan KNN dengan akurasi 87%

CONCLUSION

This study compares five different machine learning algorithms in predicting loan defaults or approvals based on consumer behavior. The XGBoost algorithm exhibits the highest accuracy, reaching 89%, followed by logistic regression and random forest with identical accuracy rates of 88%. Decision tree and KNN show slightly lower accuracies, each at 87%. Performance evaluation also encompasses precision, recall, and F1 scores, providing further insight into the prediction quality. Subsequent research can conduct experiments to optimize the parameters of each algorithm to enhance the model's performance further.

REFERENCES

[1] U. Aslam, H. I. Tariq Aziz, A. Sohail, and N. K. Batcha, “An empirical study on loan default prediction models,” *J. Comput. Theor. Nanosci.*, vol. 16, no. 8, pp. 3483–3488, 2019.

[2] C. W. Su, F. Liu, M. Qin, and T. Chnag, “Is a consumer loan a catalyst for confidence?,” *Econ. Res. Istraživanja*, vol. 36, no. 2, p. 2142260, 2023.

[3] K. Fatmawati, “GROSS DOMESTIC PRODUCT: Financing & Investment Activities and State Expenditures,” *KINERJA J. Manaj. Organ. dan Ind.*, vol. 1, no. 1, pp. 11–18, 2022.

[4] A. Gupta, V. Pant, S. Kumar, and P. K. Bansal, “Bank Loan Prediction System using Machine Learning,” in *2020 9th International Conference System Modeling and*

Advancement in Research Trends (SMART), 2020, pp. 423–426. doi: 10.1109/SMART50582.2020.9336801.

[5] G. T. H. Vuong, P. T. T. Phan, C. X. Nguyen, D. M. Nguyen, and K. D. Duong, “Liquidity creation and bank risk-taking: Evidence from a transition market,” *Heliyon*, vol. 9, no. 9, 2023.

[6] N. Muhamad, “Gen Z dan Milenial Jadi Penyumbang Kredit Macet Pinjol Terbesar pada Juni 2023,” *Databoks*, p. 2023, 2023. [Online]. Available: https://databoks.katadata.co.id/datapublish/2023/08/23/gen-z-dan-milenial-jadi-penyumbang-kredit-macet-pinjol-terbesar-pada-juni-2023

[7] S. I. Serengil, S. Imece, U. G. Tosun, E. B. Buyukbas, and B. Koroglu, “A Comparative Study of Machine Learning Approaches for Non Performing Loan Prediction,” in *2021 6th International Conference on Computer Science and Engineering (UBMK)*, 2021, pp. 326–331. doi: 10.1109/UBMK52708.2021.9558894.

[8] J. W. Goodell, S. Kumar, W. M. Lim, and D. Pattnaik, “Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis,” *J. Behav. Exp. Financ.*, vol. 32, p. 100577, 2021.

[9] D. Mhlanga, “Financial inclusion in emerging economies: The application of machine learning and artificial intelligence in credit risk assessment,” *Int. J. Financ. Stud.*, vol. 9, no. 3, p. 39, 2021.

[10] I. Lee and Y. J. Shin, “Machine learning for enterprises: Applications, algorithm selection, and challenges,” *Bus. Horiz.*, vol. 63, no. 2, pp. 157–170, 2020.

[11] A. R. Provenzano *et al.*, “Machine learning approach for credit scoring,” *arXiv Prepr. arXiv2008.01687*, 2020.

[12] S. Sobana and P. J. L. Ebenezer, “A Comparative Study of Machine Learning Algorithms for Loan Approval Prediction,” *Int. Res. J. Mod. Eng. Technol. Sci.*, vol. 04, no. 12, p. 565, 2022.

[13] T. Aditya Sai Srinivas, S. Ramasubbareddy, and K. Govinda, “Loan Default Prediction Using Machine Learning Techniques,” in *Innovations in Computer Science and Engineering: Proceedings of the Ninth ICICSE, 2021*, Springer, 2022, pp. 529–535.

[14] S. Surana, “Loan Prediction Based on Customer Behavior,” *kaggle.com*, 2021. https://www.kaggle.com/datasets/subhamjain/loan-prediction-based-on-customer-behavior/data (accessed Dec. 18, 2023).

[15] S. Maldonado, J. López, and C. Vairetti, “An alternative SMOTE oversampling strategy for high-dimensional datasets,” *Appl. Soft Comput.*, vol. 76, pp. 380–389, 2019.

[16] A. Das, “Logistic regression,” in *Encyclopedia of Quality of Life and Well-Being Research*, Springer, 2021, pp. 1–2.

[17] J. C. Stoltzfus, “Logistic regression: a brief primer,” *Acad. Emerg. Med.*, vol. 18, no. 10, pp. 1099–1104, 2011.

[18] B. Charbuty and A. Abdulazeez, “Classification based on decision tree algorithm for machine learning,” *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, pp. 20–28, 2021.

[19] I. S. Damanik, A. P. Windarto, A. Wanto, Poningsih, S. R. Andani, and W. Saputra, “Decision tree optimization in C4. 5 algorithm using genetic algorithm,” in *Journal of*

Physics: Conference Series, IOP Publishing, 2019, p. 12012.
[20] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.
[21] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
[22] J. D. Novaković, A. Veljović, S. S. Ilić, Ž. Papić, and M. Tomović, "Evaluation of classification models in machine learning," *Theory Appl. Math. Comput. Sci.*, vol. 7, no. 1, p. 39, 2017.