Journal of Information Systems Engineering and Management
2025, 10(36s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Predicting Bank Loan Risk Using Incremental Learning

Jiyan Suleiman Dahir

Pathology and Microbiology Dept., Veterinary Medicine Collage, University of Duhok, Duhok, Iraq

jeean.daher@uod.ac

ARTICLE INFO

Received: 24 Dec 2024
Revised: 16 Feb 2025
Accepted: 28 Feb 2025

ABSTRACT

Predicting loan default is a critical issue in the banking sector, given the financial risks it entails, which impact banks' performance and stability. This paper proposes incremental learning-based models for classifying bank loans and estimating their probability of default, comparing these models with the performance of traditional one-shot models. A set of machine learning models, including logistic regression, decision tree, random forest, XGBoost, SVM, and multilayer neural network (MLN), were tested on data comprising financial and behavioral information on bank customers. The results showed that the XGBoost model achieved the best performance in terms of accuracy and stability in the incremental learning model, reaching 0.9340.

Keywords: Bank loans, incremental learning, risk classification, machine learning algorithms, XGBoost.

1. Introduction

Credit risk is defined as the potential loss a bank may incur due to a credit client's failure to comply with contractual requirements and fulfill their obligations partially or fully on time [1][2][3]. Incorrect customer selection, contractual deficiencies, the client's insufficient financial strength to meet their responsibilities, assigning a credit limit that is too high to match their income/debt repayment balance, insufficient collateral received, and economic factors all contribute to the risk of loan insolvency [1][4][5]. Credit risk, therefore, is the situation in which a client fails to fully fulfill the obligations they have undertaken under the contract concluded between the bank and the client within the period specified in the contract. In other words, it is the situation in which a client is unable to repay the interest-bearing loan they have obtained from the bank on time (default) [6][7]. Credit risk is the greatest risk facing banks. As a result of poor credit risk policies that have had serious repercussions on both the financial and non-financial sectors, the importance of risk measurement has emerged, leading to an increase in research on banking risk regulations [8][9]. In this context, numerous research and analytical techniques and machine learning models have been developed to quantify credit default risk. Many different approaches have been proposed to address this issue. While techniques such as logistic regression and discriminant analysis are traditionally used to estimate credit default risk, machine learning algorithms have also emerged as a new approach, and their applications are increasing in the field of credit risk, in addition to their use in many different areas [10][11][12][13][14]. It will be critical for financial institutions to effectively manage their credit risk. Although many machine learning algorithms are used to estimate credit default risk, there is no consensus on which technique provides the best performance [15][16][17]. Therefore, the classification success of models created using algorithms is measured using various statistical and machine learning techniques, and the model that provides the best classification performance is identified as the optimal model. Bank borrowing data suffers from data imbalance, as most loans are repaid and a small percentage default, which in turn poses a challenge when training traditional machine learning algorithms [18][19][20]. Different

Journal of Information Systems Engineering and Management
2025, 10(36s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

methods can be used to address dataset imbalance during the data initialization process. Such methods are effective when models are trained statically once, and the data is available for training only once [21][22][23]. However, it is not suitable for the nature of banking data, as it is variable and continuous, and requires the development of models that can adapt to imbalance without the need for preprocessing the data in each training batch. This paper proposes the use of incremental xgboost models to model bank loan risk.

2. Related Work

Several authors have tested the use of machine learning models to predict bank loan risk and developed modified models. In [24], the authors applied three models—Gradient Boosting, XGBoost, and AdaBoost—to a dataset comprising banking and demographic information. The results showed that XGBoost performed better than other models, with an accuracy of 0.95. In [25], the researchers used auto loan data from the Kaggle platform. They applied Filter-Wrapper to feature selection and Smote-Tomek link to address data imbalance. They proposed the PSO-XGBoost model, which combines the XGBoost algorithm and particle swarm optimization. The results showed that the proposed model outperformed other models. The authors in [26] proposed using XGBoost to predict the size of social financing. The data included a range of financial indicators, including bonds and credit loans. To determine the importance of features, they used the SHAP model interpretation. The results showed that the impact of features on prediction accuracy varied. The authors in [27] proposed using XGBoost and modifying it to address imbalance in credit scoring data, applying a modified loss function and using a link function based on the extreme value (GEV) distribution. They used Freddie Mac mortgage data, which had a lower default rate. These changes led to the detection of rare defaults. In [28], the authors proposed the IAHE model to address the problems of big data and consumer behavior changes over time. The model is based on the concept of incremental learning to adapt to data variability. Its performance was evaluated and compared with nine models, and experiments were conducted on four datasets. The results demonstrated the model's ability to adapt to data fluctuations while maintaining good classification performance. In [29], instead of relying on financial indicators, the researchers developed a scoring system for behavioral and financial indicators to assess credit risk in small businesses. After feature selection using XGBoost, they used a modified random forest model. The results demonstrated the superiority of the proposed model when behavioral indicators were added to the classification. In [30], they proposed a method for finding XGBoost parameters using the particle swarm algorithm (PSO). They varied the swarm's partitioning into clusters and used different learning strategies. The model was tested on four credit datasets and seven benchmark datasets. The results showed that the proposed partitioning method was more effective than traditional methods such as grid search in finding hyperparameters. Machine learning models have proven effective in bank loan decision-making. Changing economic and financial conditions of individuals and businesses over time may affect the reliability of trained machine learning models. Studies conducted have not addressed model testing in stages rather than batch training.

3. Methodology

This paper proposes the use of incremental learning in modeling bank loans. Initially, a dataset containing information about bank loans is identified. The dataset is then initialized to compensate for missing values and convert categorical data to numeric data. The data is then divided into two sets: 80% for training and 20% for testing. Several machine learning algorithms, such as linear regression (LR), decision trees (DT), random forests (RF), gradient boosting (XGB), support vector machines (SVM), and neural networks (MLP), are tested to select the best algorithm. Finally, the performance of the models is evaluated in batches and as a whole. After selection and evaluation, a prototype is created, and the model is then gradually trained on small portions of data (batches) repeatedly. After all updates and refinements are completed, the final model is obtained, ready for use in actual predictions and

Journal of Information Systems Engineering and Management
2025, 10(36s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

evaluation of bank loan applications. The implementation proceeds as shown in the flowchart in figure 1:

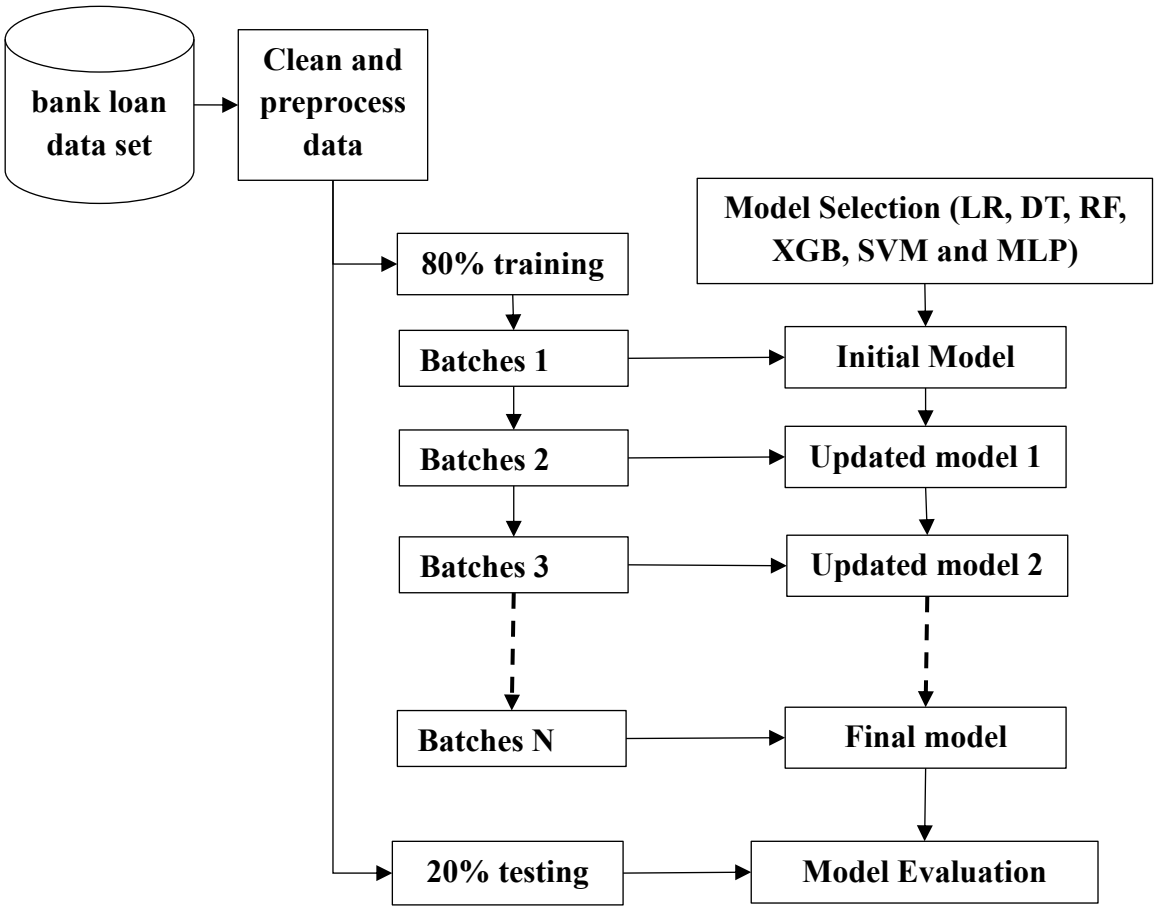

Figure 1: Proposed method for batch training models

3.1. Incremental models

In this paper, models are modified and trained in batches, with batches varying for each model.

Decision Tree: In cumulative training of a decision tree, the warm_start=True parameter is enabled with the partial_fit function. When new data arrives, the tree recalculates split points based only on the new data, while preserving the basic structure of the existing tree. The tree is not rebuilt from scratch; existing branches are gradually modified.

Random Forest: Warm_start=True is used in the RandomForestClassifier to enable cumulative training. When new data is added, the number of trees (n_estimators) gradually increases, as new trees are built on the added data while keeping the old trees unchanged. Each new tree is trained on a random portion of the updated data.

SVM (Support Vector Machine): To convert an SVM to a cumulative model, the SGDClassifier with loss='hinge' is used, which is based on the stochastic gradient descent algorithm. The model weights are updated incrementally with each new batch of data. The gradient is calculated for each sample and the model is adjusted accordingly, while preserving the most important support vectors.

In cumulative XGBoost, xgb.train() is used with the incremental update feature. New trees (boosting rounds) can be added to the current model, or existing trees can be modified using

Journal of Information Systems Engineering and Management
2025, 10(36s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

process_type='update'. Each new batch of data is used to improve the leaf weights or to build additional trees to compensate for the remaining errors.

For multi-layer neural networks (MLP), the MLPClassifier is used with the partial_fit function. The data is passed in batches, with the gradients for each batch calculated and used to adjust the neural network weights via backpropagation. The weights are updated incrementally without re-initializing the model.

To make logistic regression cumulative, the SGDClassifier is used with loss='log_loss'. This tool uses stochastic gradient regression to gradually adjust the model parameters with each batch.

Table 1 A parameters that were used to incrementally train the models.

models	parameters
Decision Tree	Retrain gradually using warm_start=True
Random Forest	warm_start=True and add progressive trees
SVM	SGDClassifier(loss='hinge')
XGBoost	Update progressively using xgb.train()
Neural Network	MLPClassifier with partial_fit
Logistic Regression	SGDClassifier(loss='log_loss')

3.2. Dataset

To conduct the experiments, a dataset from the machine learning platform Kaggle [31] was used. The dataset contains 100,515 bank loans and 19 features describing each loan and its associated customer. The features used include a range of loan information, including information about the loan itself (such as loan amount, repayment term, and purpose), customer information (such as annual income, credit score, and home ownership), credit history (such as the number of open accounts and previous credit problems), and whether the customer has repaid the loan or defaulted. Table 2 describes the dataset used.

Table 2: Dataset description

Field Name	Description	Data Type	Example
Loan ID	identifier for the loan	Text	14dd8831-6af5-...
Customer ID	identifier for the customer	Text	981165ec-3274-...
Loan Status	Status of the loan	Text	Fully Paid
Current Loan Amount	Current loan amount	Numeric	445,412
Term	Loan term	Text	Short Term
Credit Score	Customer's credit score	Numeric	709
Annual Income	Customer's yearly income	Numeric	1,167,493
Years in Current Job	Number of years in current job	Text	8 years
Home Ownership	Homeownership status	Text	Home Mortgage
Purpose	Purpose of the loan	Text	Home Improvements
Monthly Debt	Customer's monthly debt payments	Numeric	5,214.74

Journal of Information Systems Engineering and Management
2025, 10(36s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Years of Credit History	Length of credit history	Numeric	17.2
Months Since Last Delinquent	Months since last payment delinquency	Numeric	8
Number of Open Accounts	Total open credit accounts	Numeric	6
Number of Credit Problems	Past credit issues	Numeric	1
Current Credit Balance	Current outstanding credit balance	Numeric	228,190
Maximum Open Credit	Total available credit limit	Numeric	416,746
Bankruptcies	Number of past bankruptcies	Numeric	1
Tax Liens	Number of tax liens on record	Numeric	0

4. Results

To determine the effectiveness of incremental learning and compare it to static learning, we compared the models used in classifying bank loans, and conducted two different training experiments: batch training and incremental learning. The performance of each model was evaluated using four metrics: accuracy, precision, recall, and the F1 measure. The models were gradually trained on successive batches of data, and their performance was tracked during the training period. Figure 2 shows the results across batches.

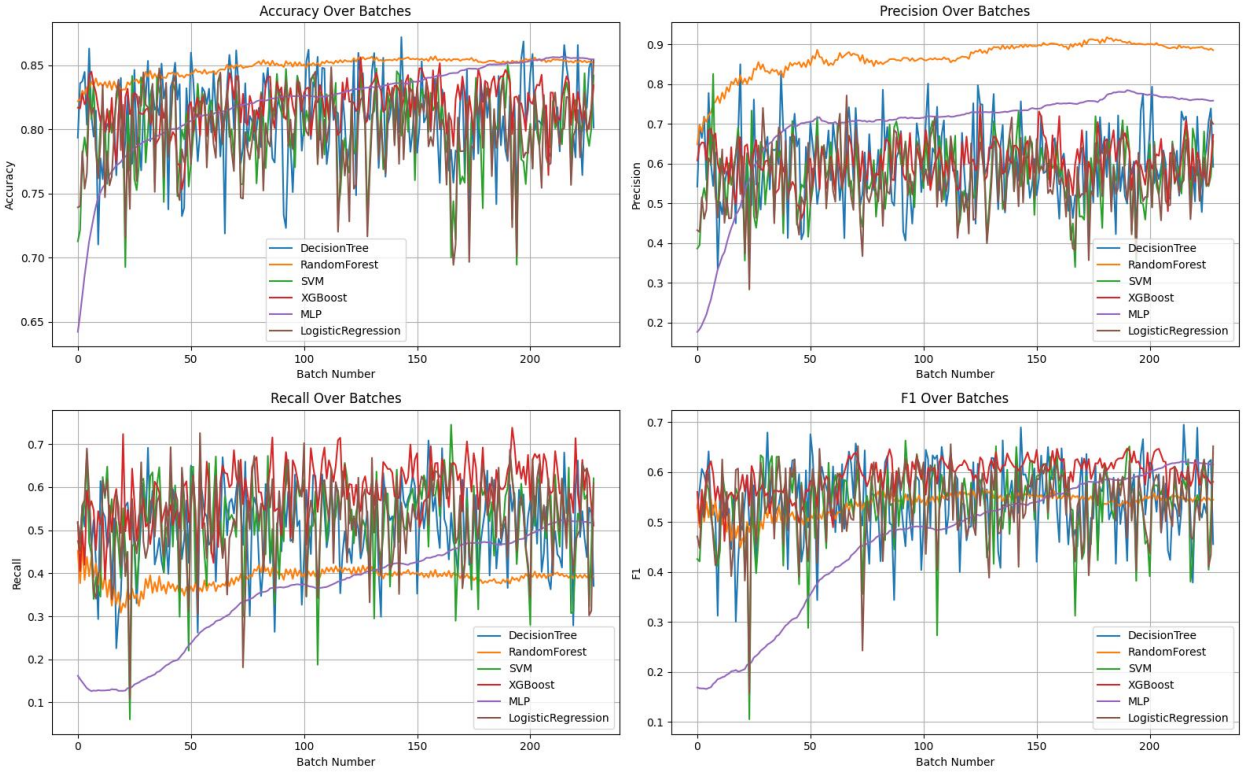

Figure 2 the performance across batches

Journal of Information Systems Engineering and Management
2025, 10(36s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Two performance tables from different testing phases are presented to measure the overall efficiency of the classifiers. Table 3 presents the results of incremental learning; the overall performance values of the models were generally higher, especially in terms of stability across scales. Table 4 shows the performance of the models when they were trained on the entire dataset at once.

Table 3. the performance of incremental learning

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	0.8631	0.8563	0.8631	0.8540
Decision Tree	0.8909	0.8921	0.8909	0.8914
Random Forest	0.9310	0.9332	0.9310	0.9273
XGBoost	0.9340	0.9357	0.9340	0.9307
SVM	0.8645	0.8579	0.8645	0.8558
MLP	0.8547	0.7589	0.5167	0.6148

Table 4. the performance of the models based one batch.

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	0.8539	0.7004	0.6093	0.6517
Decision Tree	0.8015	0.5920	0.3704	0.4557
Random Forest	0.8525	0.8860	0.3930	0.5445
SVM	0.8422	0.6570	0.6202	0.6381
XGBoost	0.8345	0.6728	0.5105	0.5805
MLP	0.8547	0.7589	0.5167	0.6148

5. Discussion

The outcomes show how machine learning models, whether trained in one batch or incrementally learning, behave when classifying bank loans.

A comparison of the two tables showed that there was a difference in the models' performance between the two training approaches. The accuracy and recall of many models decreased when trained on batch data, the incrementally trained models produced consistent and better results across the majority of metrics. This suggests that incremental learning enhances the model's capacity to adjust to evolving data, which makes it more appropriate for banking data, which frequently undergoes changes over time.

Models such as XGBoost and Random Forest performed better than the other models in both cases, but were more stable in the incremental learning case. This reflects the effectiveness of ensemble-based models in handling complex and variable data.

The decision tree model showed fluctuating performance in the batch training scenario, with recall declining, suggesting it may be sensitive to data distribution and unable to generalize well when trained on a large dataset at once.

The logistic regression and SVM models were relatively stable, maintaining acceptable levels of performance in both cases. The performance of the multilayer neural network (MLN) was uneven, gradually improving during incremental learning but not reaching high levels of recall, indicating the need for a larger number of iterations or additional parameter tuning.

One indicator that revealed clear differences between the training methods was the recall index, which reflects the model's ability to recognize positive cases (such as loan defaults). In the batch training scenario, recall declined significantly in most models, potentially leading to the omission of important defaults. However, in the incremental learning scenario, this value improved, indicating that the models were able to gradually capture patterns and better recognize these cases.

6. Conclusions

Incremental learning has been proposed for use in banking loan classification. Experimental results demonstrate that incremental learning is more efficient when dealing with variable data such as bank customer data. Unlike batch training, it helped improve model accuracy over time, especially in metrics such as recall, which is important for detecting positive cases such as non-performing loans. On the other hand, ensemble models such as XGBoost and Random Forest demonstrated more stable performance, both in terms of accuracy and balance across the four metrics, indicating their adaptability to new data batches. Other models, such as decision trees and logistic regression, performed less consistently, especially when trained in batches. Furthermore, the MLP neural network required more batches to begin improving, indicating that some models require more time and fine-tuning to improve performance, especially in variable data scenarios. Future work could improve the handling of data imbalance within incremental learning, and model testing could be done on live data to monitor its performance in real time. Automatic parameter tuning techniques could also be used to improve model accuracy.

References

[1] Z. Wang, "Artificial intelligence and machine learning in credit risk assessment: Enhancing accuracy and ensuring fairness," *Open Journal of Social Sciences*, vol. 12, no. 11, pp. 19–34, Jan. 2024, doi: 10.4236/jss.2024.1211002.

[2] R. D. Modi and P. D. Tawde, "Credit score prediction using machine learning," *Int. J. Adv. Res. Sci. Commun. Technol.*, vol. 10, no. 2, pp. 145–149, Feb. 2025, doi: 10.48175/IJARSCT-23348.

[3] E. Ileberi, Y. Sun, and Z. Wang, "A machine learning-based credit risk prediction engine system using a stacked classifier and a filter-based feature selection method," *J. Big Data*, vol. 11, no. 1, pp. 1–15, Feb. 2024, doi: 10.1186/s40537-024-00882-0.

[4] R. M. Darst, E. Refayet, and A. Vardoulakis, "Banks, non-banks, and lending standards," *Finance and Economics Discussion Series*, no. 86, pp. 1–52, Oct. 2020, doi: 10.17016/feds.2020.086.

[5] V. Chang et al., "Credit risk prediction using machine learning and deep learning: A study on credit card customers," *Risks*, vol. 12, no. 11, p. 174, Nov. 2024, doi: 10.3390/risks12110174.

[6] M. Naili and Y. Lahrichi, "Banks' credit risk, systematic determinants and specific factors: Recent evidence from emerging markets," *Heliyon*, vol. 8, no. 2, Feb. 2022, Art. no. e08960, doi: 10.1016/j.heliyon.2022.e08960.

[7] R. Chhetri, R. Parajuli, and G. Sharma, "Credit risk prediction by using ensemble machine learning algorithms," *Int. J. Res. Publ.*, vol. 147, no. 1, Apr. 2024, doi: 10.47119/IJRP1001471420246296.

[8] S. Solanki and A. Professor, "The effectiveness of different credit risk assessment methods on loan performance," *J. Cardiovasc. Dis. Res.*, vol. 12, no. 6, Jun. 2023, doi: 10.48047/jcdr.2021.12.06.335.

[9] A. O. Ikudabo and P. Kumar, "AI-driven risk assessment and management in banking: Balancing innovation and security," *Int. J. Res. Publ. Rev.*, vol. 5, no. 10, pp. 3573–3588, Oct. 2024, doi: 10.55248/gengpi.5.1024.2926.

[10] G. M. Jakka et al., "A novel credit scoring system in financial institutions using artificial intelligence technology," *J. Auton. Intell.*, vol. 6, no. 2, p. 824, Aug. 2023, doi: 10.32629/jai.v6i2.824.

[11] J. P. Noriega, L. Rivera, and J. Herrera, "Machine learning for credit risk prediction: A systematic literature review," *Data*, vol. 8, no. 11, p. 169, Nov. 2023, doi: 10.3390/data8110169.

[12] M. Sun et al., "Applying hybrid graph neural networks to strengthen credit risk analysis," *arXiv preprint arXiv:2410.04283*, Oct. 2024. [Online]. Available: https://arxiv.org/abs/2410.04283

[13] P. Kündig and F. Sigrist, "A spatio-temporal machine learning model for mortgage credit risk: Default probabilities and loan portfolios," *arXiv preprint arXiv:2410.02846*, Oct. 2024. [Online]. Available: https://arxiv.org/abs/2410.02846

[14] W. A. Addy et al., "AI in credit scoring: A comprehensive review of models and predictive analytics," *Glob. J. Eng. Technol. Adv.*, vol. 18, no. 2, Feb. 2024, doi: 10.30574/gjeta.2024.18.2.0029.

[15] J. Alvi, I. Arif, and K. Nizam, "Advancing financial resilience: A systematic review of default prediction models and future directions in credit risk management," *Heliyon*, vol. 10, no. 21, Oct. 2024, doi: 10.1016/j.heliyon.2024.e39770.

[16] I. E. Ahmed, R. Mehdi, and E. A. Mohamed, "The role of artificial intelligence in developing a banking risk index: An application of adaptive neural network-based fuzzy inference system (ANFIS)," *Artif. Intell. Rev.*, vol. 56, no. 11, pp. 13873–13895, Apr. 2023, doi: 10.1007/s10462-023-10473-9.

[17] B. E. Abikoye and C. Agorbia-Atta, "How artificial intelligence and machine learning are transforming credit risk prediction in the financial sector," *Int. J. Sci. Res. (IJSR)*, vol. 12, no. 2, pp. 1884–1889, Aug. 2024, doi: 10.30574/ijsra.2024.12.2.1467.

[18] J. Liu, X. Zhang, and H. Xiong, "Credit risk prediction based on causal machine learning: Bayesian network learning, default inference, and interpretation," *J. Forecast.*, vol. 43, no. 5, pp. 1234–1256, Feb. 2024, doi: 10.1002/for.3080.

[19] N. S. Alfaiz and S. M. Fati, "Enhanced credit card fraud detection model using machine learning," *Electronics*, vol. 11, no. 4, p. 662, Feb. 2022, doi: 10.3390/electronics11040662.

[20] T. Mokheleli and T. Museba, "Machine learning approach for credit score predictions," *J. Inf. Syst. Inform.*, vol. 5, no. 2, pp. 497–517, May 2023, doi: 10.51519/journalisi.v5i2.487.

[21] A. Akinjole et al., "Ensemble-based machine learning algorithm for loan default risk prediction," *Mathematics*, vol. 12, no. 21, p. 3423, Oct. 2024, doi: 10.3390/math12213423.

[22] Y. Huang et al., "Uncertainty-aware learning against label noise on imbalanced datasets," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 6, pp. 6734–6742, Jun. 2022, doi: 10.1609/aaai.v36i6.20654.

[23] S. Galli, "Dealing with imbalanced datasets in machine learning: Techniques and best practices," *Train in Data*, Mar. 2023. [Online]. Available: https://www.blog.trainindata.com/machine-learning-with-imbalanced-data/

[24] C. Rao, Y. Liu, and M. Goh, "Enhancing financial decision-making: Predictive modeling for personal loan eligibility with gradient boosting, XGBoost, and AdaBoost," *Inf. Technol. Econ. Bus.*, vol. 1, no. 1, pp. 7–13, 2024, doi: 10.69882/adba.iteb.2024072.

[25] C. Rao, Y. Liu, and M. Goh, "Credit risk assessment mechanism of personal auto loan based on PSO-XGBoost model," *Complex Intell. Syst.*, vol. 9, pp. 1391–1414, 2023, doi: 10.1007/s40747-022-00854-y.

[26] J. Zuo, C. Bao, Q. Meng, and Q. Zheng, "A study on the incremental size of social financing based on XGBoost and SHAP," *Procedia Comput. Sci.*, vol. 221, pp. 1321–1328, 2023, doi: 10.1016/j.procs.2023.08.121.

[27] J. Mushava and M. Murray, "A novel XGBoost extension for credit scoring class-imbalanced data combining a generalized extreme value link and a modified focal loss function," *Expert Syst. Appl.*, vol. 202, p. 117233, 2022, doi: 10.1016/j.eswa.2022.117233.

[28] T. Museba, "Incremental machine learning-based approach for credit scoring in the age of big data," in *Towards Digitally Transforming Accounting and Business Processes*, T. Moloi and B. George, Eds., Cham, Switzerland: Springer, 2024, pp. ICAB 2023, doi: 10.1007/978-3-031-46177-4_29.

[29] F. Fang, "Credit risk evaluation model of small-micro enterprises for rural commercial bank based on XGBoost and random forest," in *Proc. 2022 Int. Conf. E-business and Mobile Commerce (ICEMC '22)*, pp. 125–131, 2022, doi: 10.1145/3543106.3543127.

[30] C. Qin et al., "XGBoost optimized by adaptive particle swarm optimization for credit scoring," *J. Adv. Comput. Intell. Intell. Informatics*, vol. 2021, p. 6655510, 2021, doi: 10.1155/2021/6655510.

[31] Z. Begiev, "My Dataset," *Kaggle*, 2022. [Online]. Available: https://www.kaggle.com/datasets/zaurbegiev/my-dataset