December 2025 | Volume 1 | Issue 3 | Article No. 25215

GR JOURNALS

Journal of Smart Sensors and Computing

GR

Journal of Smart Sensors and Computing

Research Article | *Open Access* |

Loan Default Prediction Using Ensemble Machine Learning Algorithms

Sanjay Gour* and Pooja Soni

Department of Computer Science & Engineering, Gandhinagar University, Gandhinagar, 382725, Gujarat, India
**Email:* sanjay.since@gmail.com (S. Gour)

Abstract

Loan default prediction has become a critical task for organizations operating in the financial sector, as it directly influences risk management, loan approval decisions, and overall organizational profitability. Traditional credit assessment methods employed by financial institutions rely on a limited set of predefined factors and often fail to effectively capture complex patterns associated with loan default behavior. Consequently, these approaches are insufficient for accurately identifying potential defaulters, leading to increased financial risk. To address these limitations, this study focuses on evaluating the performance of several ensemble machine learning algorithms, including Random Forest, Gradient Boosting, XGBoost, and LightGBM, for loan default prediction. An experimental methodology is adopted using a publicly available benchmark dataset. The workflow involves data preprocessing, feature engineering, class imbalance handling, model training, and performance evaluation. The effectiveness of the proposed models is assessed using standard evaluation metrics such as accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (ROC-AUC). In addition, a detailed analysis based on the confusion matrix is conducted to examine classification performance. The results demonstrate the strong capability of ensemble machine learning techniques in accurately predicting loan defaults and highlight their effectiveness in feature-driven predictive modeling within the financial domain.

***Keywords*:** Machine learning; Loan default; Algorithms; Random Forest; XGBoost; LightGBM.

Received: 11 November 2025; Revised: 26 December 2025; Accepted: 30 December 2025; Published Online: 31 December 2025.

1. Introduction

A loan default refers to the failure of a borrower to fulfill the agreed repayment obligations within the stipulated schedule, either partially or in full.[1] Defaults may occur due to various factors, including financial instability, unemployment, excessive debt burden, unexpected economic conditions, or poor credit behavior. From the lender's perspective, loan defaults represent a significant source of financial loss and increased operational risk, as they directly impact asset quality, liquidity, and profitability.[2] Consequently, accurately identifying high-risk borrowers prior to loan approval has become a critical requirement for financial institutions.[3] Early detection of potential loan defaulters enables proactive risk management, improved credit allocation, and the implementation of preventive strategies such as adjusted interest rates, collateral requirements, or alternative repayment plans. With the growing availability of large-scale financial data, data-driven approaches, particularly machine learning techniques-offer a promising solution for modeling complex borrower behavior and improving the accuracy of loan default prediction.[3-5] Machine learning (ML), a subfield of artificial intelligence, has the capability to efficiently handle large volumes of data and extract meaningful patterns from complex datasets. Algorithms belonging to the same or similar domains present two major assessment challenges. The first challenge

DOI: https://doi.org/10.64189/ssc.25215

involves comparative evaluation among algorithms applied to the same dataset within a single study, while the second challenge arises when comparing outcomes reported by different authors across related studies. The present study addresses both evaluation perspectives, as discussed in prior works.[6,7]

This study adopts a classification-based approach using decision tree–based ensemble models to support accurate and timely prediction of borrowers' likelihood of loan default. Loan default prediction is widely recognized as a critical and risk-sensitive task for financial institutions, as it directly influences lending decisions and risk mitigation strategies. In the context of growing societal dependence on data-driven decision-making, effective network-based data analysis plays a crucial role in minimizing activities that adversely affect socio-economic growth.[8]\

The motivation for this work stems from the rapid advancement of data analytics, which has significantly transformed financial institutions, particularly in credit risk assessment and loan default prediction. Traditional approaches for handling loan defaults were primarily based on statistical and data mining techniques such as logistic regression, credit scoring models, and rule-based decision systems. While these methods are effective for modeling linear relationships, they exhibit limited capability in capturing complex non-linear patterns present in large-scale financial datasets. Predicting loan defaults in such datasets is a challenging task due to the presence of multiple interacting non-linear features, a concern relevant across various industries, including FMCG and automation sectors.[9,10]

A loan default typically occurs when a borrower fails to meet scheduled repayment obligations, resulting in financial distress for the lender. The ability to identify potential defaults at an early stage enables financial institutions to proactively manage and mitigate future risks. Consequently, evaluating and assessing the performance of machine learning algorithms in this domain has become a significant area of both academic research and practical application. Performance assessment focuses on determining how effectively models can distinguish between potential defaulters and reliable borrowers.

\The evaluation process relies on quantitative performance metrics, including accuracy, precision, recall, F1-score, receiver operating characteristic (ROC) curve, and confusion matrix analysis. A comparative assessment using multiple metrics is essential to ensure a robust and unbiased evaluation of predictive performance. Furthermore, the outcomes of this study are compared with results reported in related works to establish consistency and reliability. Several machine learning algorithms have demonstrated strong performance in loan default prediction, notably Random Forest, Gradient Boosting Machines, XGBoost, and LightGBM.[11,12] These ensemble models exhibit superior performance due to their ability to capture complex, multi-level interactions among financial attributes such as credit history, income stability, debt-to-income ratio, and loan characteristics.

2. Literature review

Rendering to the research since 2019 to 2025, ensemble or collaborative machine learning approaches mainly Gradient Boosting, Random Forest and XGBoost, LightGBM, consistently outstrip the classical statistical technique methods for loan default prediction. Reviews in similar vicinities which supports the study are as follows:

Chen *et al.*[13] anticipated prejudiced logistic regression with L2 penalty dataset and TF-IDF features on Chinese credit data. The refining of imbalance dataset gives positive analytics prospects as increases accuracy while reducing overfitting. Kinjole *et al.*[14] utilised the dataset of LendingClub and processes with models SVM, Random Forest, XGBoost, and ADABoost, imposing SMOTE variants to harmonize the data. SMOTE+ENNs with XGBoost attained 90.49% accuracy, whereas ensemble assembling raised it to 93.7%, display that well-adjusted data and ensembles expand the forecasting. Zhua *et al.*[15] analysed with the Lending Club dataset by utilizing Random Forest, SVM, and Logistic Regression algorithm. The Random Forest algorithms performed best in association of SMOTE improved class balance and model consistency. Leticia Monje *et al.*[16] implemented XGBoost algorithm with a proxy and fuzzy philological model on P2P loans (2007–2020), attaining high accuracy and extra interpretability for officials and initial default exposure. Luca Barbaglia et al.[17] they work on 12M European mortgages data and finding that XGBoost outstripped logistic regression. The variables Interest rate, LTV, and local economy were considered as the key predictors, prominence provincial risk variations. Mona Aly SharafEldin *et al.*[18] utilised the Egyptian bank loans dataset along with the Decision Tree, Random Forest and Gradient Boosting algorithms. It is noted that Decision Tree (Acc. 88%) achieved unsurpassed. The key forecasters variables were balance, due amount, and delinquency. Zhang *et al.*[19] assessed the algorithm XGBoost, Gradient Boosting, and LightGBM with dataset of institutional loan data. The Gradient Boosting accomplished best accuracy (0.8887), whereas XGBoost had uppermost ROC-AUC (0.9714). The study presented cost-sensitive threshold fine-tuning for directive. Kang et al.[20] (2025) considered the Kaggle loan data (148k records) and process the same with Random Forest, XGBoost and LightGBM models, utilizing SMOTE for balance the dataset. It is observed that LightGBM achieved best (Acc 0.9764, Prec 0.9747, Rec 0.9503). The significant features remained interest rate and credit type in association of target variable

3. Objectives

- The objective of the present study contains three key directions:
- To accomplice experimental with selected ensemble

machine learning models including Random Forest, Gradient boosting, XGBoost and LightGBM) for forecasting loan defaults.

- To evaluate the performance of projected model by using appropriate metrics including accuracy, precision, recall, F1-score, and ROC AUC.
- Use confusion matrix to assess the major performance of the models.

4. Hypothesis

The supposition for the experimental outlined as to compare the implication of machine learning models for better performance as:

H1: The machine learning algorithms are significantly performing to deal the prediction of the loan defaults.

5. Methodology

Fig. 1 Illustrate the research methodology. The experimental methodology of the research includes five main segments. The first step is concern from the assortment of the proper dataset. It is necessity of dataset that appropriate credit and demographic data should be available. The second stage is concern from the Data Pre-processing; the key deliberation is treatment of missing values, data encoding, normalization and dealing the class imbalances. Feature engineering and selection with REF, correlation analysis and valuations of tree analysis. The machine training process will be complete with 80:20 ratio. At this time 80% dataset is used to train the machine and 20% of the dataset endure for the testing. Subsequently train the machine several machine learning algorithm / models are used for the experimental. Here we are considering Random Forest, Gradient Boosting Machine, XGBoost and LightGBM model od machine learning for the prediction of loan defaults. The performance of the model is assessed by metrics of Confusion matrix, accuracy, precision, recall, F1-score, ROC AUC and feature performance.

5.1 Dataset

The dataset consider for the study is taken from Kaggle[21] which is accessible from the web link https: www.kaggle.com/datasets/taweilo/loan approval-classification-data, retrieved on 10 October 2025. It comprises about 45,000 records and 14 variables, which includes both numerical and categorical features. It comprises client and loan-related features which are related to forecasting loan defaults.

Table 1: Key variables used for modelling.

Feature Name	Description
Numerical Features	
person_age	Age of the borrower.
person_income	Annual income of the borrower.
loan_amnt	The amount of money requested for the loan.
loan_int_rate	The interest rate assigned to the loan. A higher rate often signifies higher perceived risk.
debt_to_income	The ratio of the borrower's total debt to their gross income, a key indicator of financial health.
credit_score	A numerical value representing the borrower's creditworthiness.
Categorical Features	
person_home_ownership	The borrower's homeownership status, with possible values like RENT, OWN, MORTGAGE.
loan_intent	The stated purpose for the loan, such as DEBTCONSOLIDATION, HOMEIMPROVEMENT, etc.
previous_loan_defaults_on_file	A binary feature indicating if the borrower has defaulted on a loan previously.

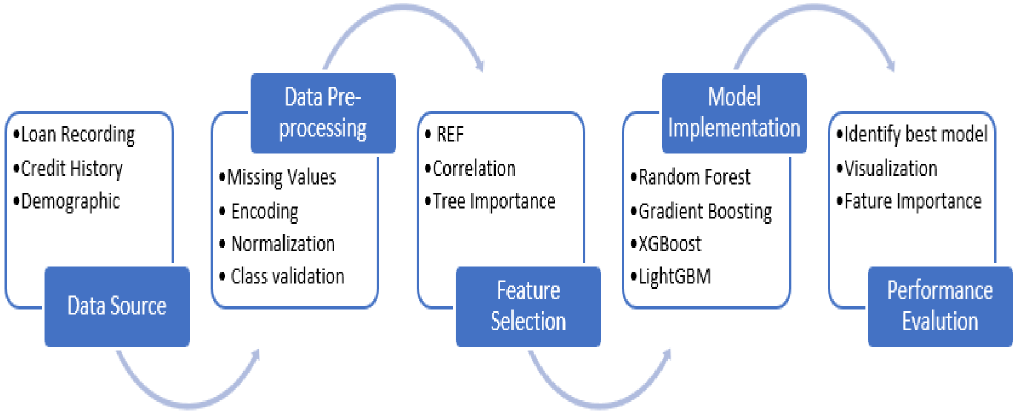

Fig. 1: Research methodology.

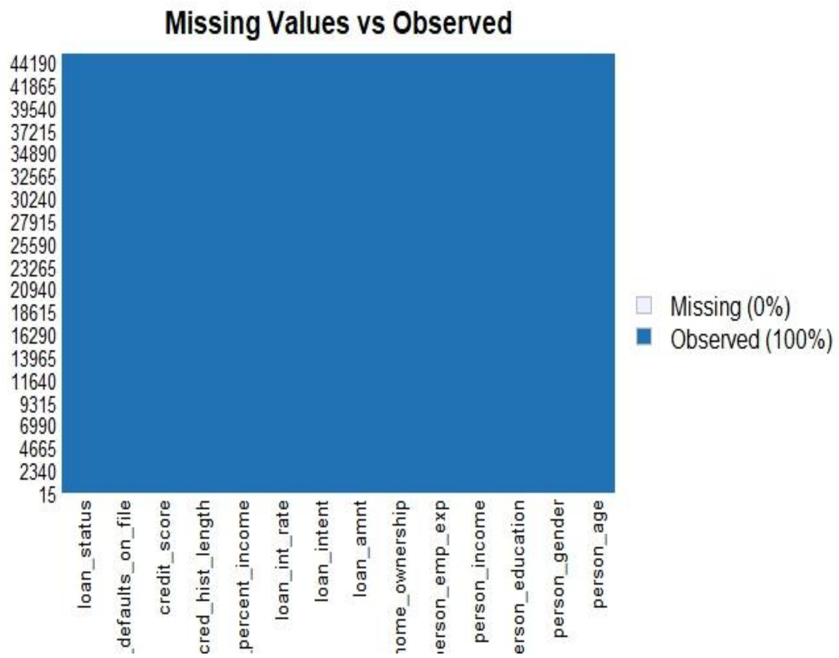

Fig. 2: Handling missing values.

5.2 Data preprocessing

The dataset was foremost examined for missing values, inconsistencies, and also for data types. Categorical attribute like as gender, education, home ownership, and loan intent remained transformed into factors and encoded by utilizing one-hot encoding process. The numeric characteristics, with loan amount, income and credit score were standardized to confirm unvarying scaling. The outliers' vales were inspected and handled suitably. Meanwhile the dataset showed class imbalance, the ROSE method was imposed to make a well-adjusted sample.[22]

5.3 Missing values

The missing value valuation was conducted to confirm data comprehensiveness and trustworthiness. In the current dataset no missing values are observed, thus dataset has no disturbance beside the missing values.[23,24]

5.4 Categorical encoding in R

There is need of transform non-numeric features into numerical feature to execute machine learning algorithms. In R language, this procedure commences by altering these attributes into factors to confirm precise credentials of categorical data. The fastDummies package is utilised to execute one-hot encoding, which generates novel binary attribute for every category inside a variable.

5.5 Normalization of numeric features

In R language, normalization might be accomplished by utilizing the scale() function, this homogenizes the data by altering every numeric attribute to have a mean of zero and a standard deviation of one. The procedure guarantees that the entire features are on a similar scale and enhance convergence for algorithms alike Random Forest, Logistic Regression, Gradient Boosting and neural networks.

5.6 Handling class imbalance

In order to tackle class imbalance, resampling methos like over-sampling, under-sampling, or a hybrid method might be utilised. In R language, the ROSE (Random Over-Sampling Examples) package delivers an effective technique for harmonizing datasets by producing artificial examples of the smaller class via random sampling and interpolation.

5.7 Feature engineering / selection

Intended for the dataset "Loan Approval Classification", the feature engineering includes forming, altering, and picking attributes that well capture outlines impacting loan approval consequences. The dataset comprises attributes like as demographics, financial attributes, and loan characteristics. In R language, this procedure commences with discovering associations between numeric attributes and their associations with the target variable which is loan_status. The derivate attribute might be formed to prompt meaningful associations.[22] Feature Selection: The feature selection is a vigorous phase in improving machine learning presentation by recognizing the utmost noteworthy predictors whereas dropping redundancy and overfitting. For the dataset Loan Approval Classification, the feature selection assistances regulate which attribute is utmost sturdily impacts the target attribute which is loan_status, confirming a extra explainable and effectual model.

5.8 Training of machine

In the present study, in order to train the machine / for learning model the ratio of 80:20 of the dataset was considered, means that first parts 80% of dataset is taken for training and 20% for testing as shown in Fig. 3.

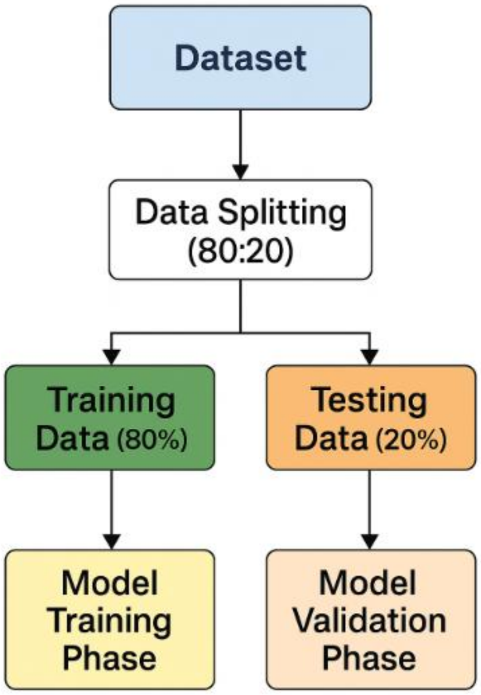

Fig. 3: Model training.

In the machine learning approach training data is utilised to crate and fitting the model, permitting it to learn outlines and associations amid attributes. The rest of 20% is kept for testing, which assesses how sound the trained model executes on unnoticed data.[25]

6. Machine learning models in R

In the projected study we are utilizing four significant model which are following the ensemble approach of machine learning.

6.1 Random Forest algorithm

The Random Forest algorithm is extensively utilised ensemble learning systems in machine learning, mainly effective for together regression and classification difficulties. It is founded on the opinion of uniting manifold decision trees to expand projecting correctness and manage overfitting.

In Random Forest, a huge amount of decision trees is constructed throughout training, and every tree harvests its individual class forecast. The closing production of the model is resolute by widely held voting (for classification tasks) or be around (for regression tasks). The main impression behindhand Random Forest is the overview of arbitrariness it randomly chooses subgroups of data (rows) and subgroups of attributes (columns) for the construction of every tree. Such kind of randomness guarantees that trees are decorrelated, thus refining the model's oversimplification competence and dropping variance.

6.2 Gradient Boosting algorithm

The Gradient Boosting Algorithm is a extremely operative ensemble learning method castoff mutually for correlation and regression work. It functions by uniting multiple puny learners characteristically decision trees into a sole robust prophetic model. Just not as bagging approaches like as Random Forest, while entire trees are constructed self-sufficiently and in equivalent mode, Gradient Boosting creates trees successively, with apiece novel tree endeavoring to accurate the residual errors thru the preceding ones. This consecutive approach permits the algorithm to gradually diminish the forecasting mistake and attain high correctness. The main knowledge behind Gradient Boosting is to enhance a loss function (like mean squared error aimed at regression or deviance for classification) utilizing gradient descent.

6.3 XGBoost algorithm

The Extreme Gradient Boosting (XGBoost) procedure is a progressive employment of the Gradient Boosting context, intended for greater speediness, efficacy, and extrapolative presentation. Established by Tianqi Chen in 2016, XGBoost has converted as widespread machine learning systems, mainly for organized and tabular statistics. It has increased extensive acceptance due to its capability to grip big datasets, avoid overfitting, and attain state-of-the-art concert together for regression and classification works. XGBoost is grounded on the belief of boosting, while an ensemble of puny learners characteristically decision trees is constructed successively. Every tree keep objective to diminish the residual mistakes created by the ensemble of beforehand trained trees. Though, dissimilar standard of Gradient Boosting, XGBoost presents numerous important improvements that brand it quicker and further vigorous.

6.4 LightGBM algorithm

The Light Gradient Boosting Machine (LightGBM) is an extremely competent and ascendable machine learning procedure established by Microsoft. It is an enactment of the Gradient Boosting Decision Tree (GBDT) outline, intended to deliver rapid training speediness, lesser memory ingesting, and well correctness, chiefly for large-scale and high-dimensional datasets. In R, LightGBM is obtainable by the package LightGBM. which permits handlers to execute classification, regression, and ranking jobs competently.

LightGBM performs by construct an ensemble of puny learners, characteristically decision trees, in a consecutive way. Every novel tree is trained to accurate the errors produced by the preceding ensemble of trees by diminishing a stated loss function. Dissimilar to old-style gradient boosting, LightGBM usages a leaf-wise growing tactic in its place of level-wise development. In such method, the algorithm cultivates the tree by excruciating the leaf through the uppermost loss lessening, follow-on in profounder trees and enhanced accurateness. Though, to avoid overfitting, the limitation max_depth might be established to bound tree deepness.

7. Tools and technologies (R Studio)

The entire study is accomplished with R language; it is an open-source platform which is mainly uses for data analysis and statistical computation. It is projected to manage data input, processing, and visualization proficiently. The R structure is alienated into three main mechanisms: 1) R Kernel, 2) R Environment, and 3) R Packages.

The IDE of R is known as the R studio, is an environment which implement the capabilities of R language at the single interface. It gives a user-friendly environment and interface with facilities of coding, reporting and visualization. Also, the library and package are comprising according to machine learning algorithm implemented in R. In R Random Forest uses "randomForest" and "ranger", for Gradient Boosting "gbm" and "caret", for XGBoost library is "xgboost" and for LightGBM library is "lightgbm".

8. Performance evaluation metrics

In order to evaluate the performance of models / algorithms validation metrics is utilised which is an organized tool to measure machine learning algorithms/ model. It characteristically comprises metrics alike accuracy, precision, recall, F1-score, and ROC-AUC. These tools are providing comparison between predicted values and actual consequences. These assistances to recognize merits and demerits, and extents for model enhancement or fine-tuning. Confusion Matrix: The Confusion Matrix is basically a squared table that displays the totals of false v/s true values classifications with actual and predicted form. It also provides a thorough breakdown of model performance. Fig. 4 is depicting binary classification (2 × 2) matrix arrangement.

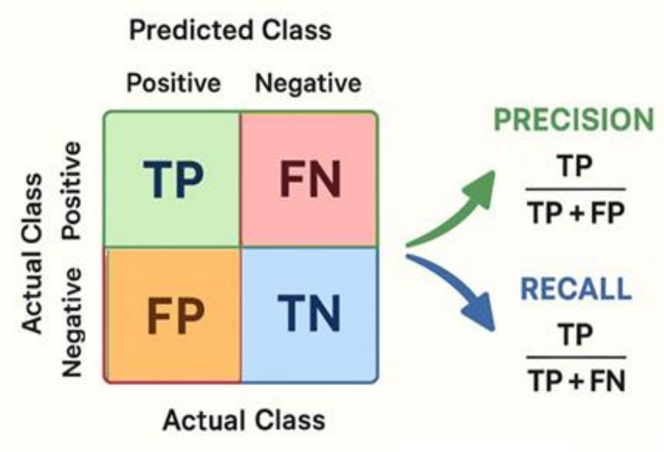

Fig. 4: Structure of confusion metrics.

The key advantage of the confusion matrix is it gives inside of values with types of error like False positive (FP) and False negative (FN). The confusion gives base to other performance measure indicators like Fi-score, precision and recall.

Accuracy: It is basically the proportion of appropriately predicted cases to the total cases. It measures in what way frequently the model is accurate all-inclusive.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Predictions}} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

F1 Score: It is the harmonic mean of Precision and Recall, which equilibriums the trade-off among them.

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}} \quad (2)$$

Precision: It is also known as the positive prophetic value. It measures the ratio of accurately predicted positive cases out of entire cases forecast as positive.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

Recall: It is also known as sensitivity or true positive rate. It assesses the ability of model to accurately recognize all positive cases.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (4)$$

ROC Curve: It is denoted as "receiver operating characteristics" curve, designs the true Positive Rate "Recall" in contradiction of the False Positive Rate "FPR" at diverse classification beginnings.

$$\text{FPR} = \frac{FP}{FP+TN} \quad (5)$$

In the plot of ROC curve the x-axis represents the FPR and y-axis represents TPR / Recall.

ROC AUC: It as denoted as Area Under ROC Curve, is a solo value brief of the ROC curve. The values are ranged between 0 to 1, where 0.5 → random guess, 1 → perfect classifier and situation <0.5 → worse than random.

Precision-Recall (PR) Curve: this curve plots precision v/s recall at diverse beginnings. The formulas of both the values are as: Precision: $\text{TP}/(\text{TP}+\text{FP})$, Recall: $\text{TP}/(\text{TP}+\text{FN})$

9. Results

The confusion metrics is utilised to create ground for the measurement of performance of the machine learning models. The models are assessed on the testing dataset and

Table 2: Description of values of confusion metrics.

Cell	Actual	Predicated	Interpretation
TN	0 (Negative)	0 (Negative)	The model correctly predicted the negative class.
FP	0 (Negative)	1 (Positive)	The model incorrectly predicted the positive class when the actual class was negative (Type I error).
FN	1 (Positive)	0 (Negative)	The model incorrectly predicted the negative class when the actual class was positive (Type II error).
TP	1 (Positive)	1 (Positive)	The model correctly predicted the positive class.

are presented below to deliver a comprehensive detail of their classification performance. The matrices are vital for analysis of classification errors, exactly particularization of the number of true positives, true negatives, false positives, and false negatives. This is vital to know how well every categorizes real defaulters (true positives) whereas lessening improper classifications of non-defaulters as defaulters (false positives) and contrariwise, which influences financial risk valuation. The numeric zero is denoted as Loan not defaulted and one as Loan defaulted

9.1 Confusion matrix for Random Forest Model

According to the values of is TN = 6820 which accurately predicted non-defaulters, FP = 179 non-defaulters forecast as defaulters, FN = 455 defaulters forecast as non-defaulters and TP = 1546 accurately forecast defaulters (Fig. 5).

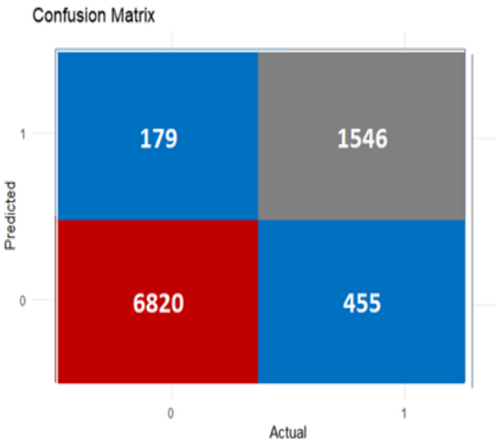

Fig. 5: Confusion matrix for Random Forest algorithm.

The accuracy is approximately 93% which shows that model is accurate for maximum predictions, but it might be misleading when dataset is imbalanced. The precision approximately 89.6% which shows that maximum people forecast as defaulters are really defaulters. Thus now 10% of forecasted defaulters are false positives. The recall approximately 77.3%, shows model accurately recognizes approximately 77% of real defaulters thus 23% of defaulters are misclassified as non-defaulters, (FN = 455). False Negatives (455) means these are actual defaulters which predicted as non-defaulters are "risk to the bank". False Positives (179) means these are non-defaulters forecasted as defaulters, possible lost business or disallowed loan applications.

9.2 Confusion matrix for Gradient Boosting Machine

The values received from confusion matrix are as the values of is TN = 6800 which accurately predicted non-defaulters, FP = 200 non-defaulters forecast as defaulters, FN = 463 defaulters forecast as non-defaulters and TP = 1537 accurately forecast defaulters (Fig. 6).

The Gradient Boosting Machine model achieves an accuracy of approximately 92.6%, indicating strong overall predictive performance. However, due to class imbalance, accuracy alone may be misleading. The precision of 88.5% suggests that most customers predicted as defaulters are indeed defaulters, with around 11% false positives, which may result in lost business opportunities or rejected loan applications. The recall of 76.8% indicates that the model correctly identifies nearly 77% of actual defaulters, while 23% (FN = 463) are misclassified as non-defaulters, posing a potential financial risk to the bank. Additionally, 200 false positives represent non-defaulters incorrectly classified as defaulters.

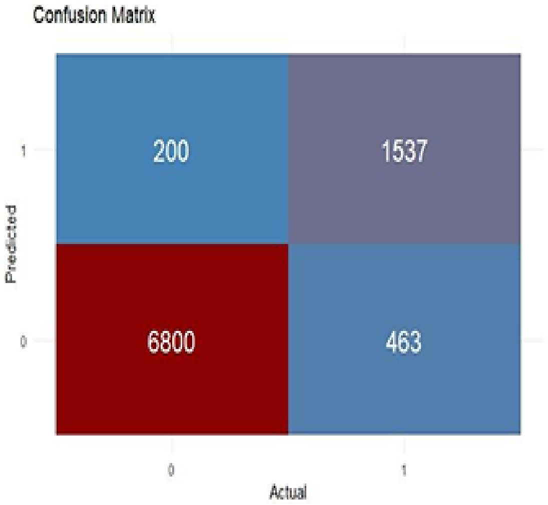

Fig. 6: Confusion matrix for Gradient Boosting machine algorithm.

9.3 Confusion matrix for XGBoost model

The values received from XGBoost are as the values of is TN = 6805 which accurately predicted non-defaulters, FP = 195 non-defaulters forecast as defaulters, FN = 421 defaulters forecast as non-defaulters and TP = 1579 accurately forecast defaulters (Fig. 7).

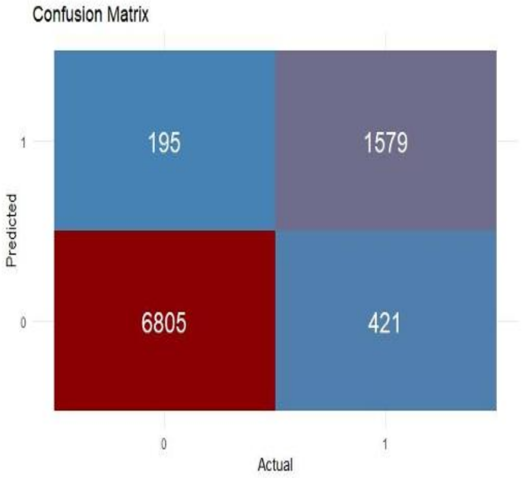

Fig. 7: Confusion matrix for XGBoost algorithm.

The model achieves an accuracy of approximately 93.2%, indicating strong overall predictive performance. The precision of about approximately suggests that most customers predicted as defaulters are indeed defaulters, with nearly 11% false positives, which may result in lost business opportunities or rejected loan applications. The recall of approximately 79% model accurately recognizes approximately 79% of real defaulters so in this case 21% of defaulters are misclassified as non-defaulters (FN = 421). False Negatives (421) means these are actual defaulters which predicted as non-defaulters which considered as "risk to the bank". False Positives (195) means these are non-defaulters forecasted as defaulters, possible lost business or disallowed loan applications.

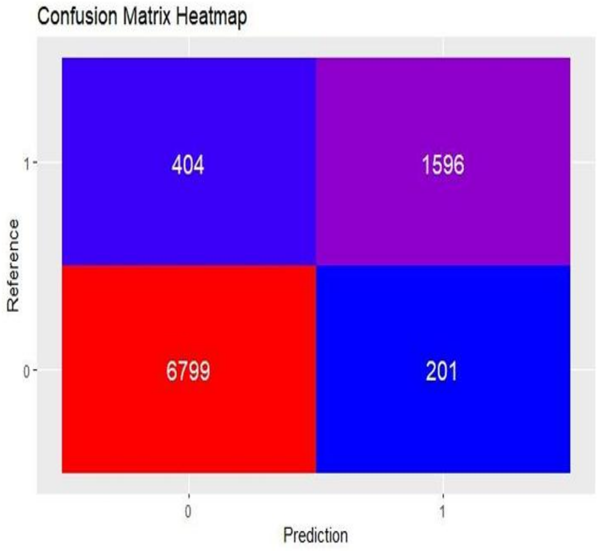

Fig. 8: Confusion matrix for LightGBM Algorithm.

9.4 Confusion matrix for LightGBM Model

The values received from the LightGBM algorithms shows that the values of is TN = 6799 which accurately predicted non-defaulters, FP = 201, non-defaulters forecast as defaulters, FN = 404 defaulters forecast as non-defaulters and TP = 1596 accurately forecast defaulters (Fig. 8).

The accuracy is approximately 93.3% which shows that model is accurate for maximum predictions. The precision approximately 88.8% shows that maximum people forecast as defaulters are really defaulters. Thus now 11% of forecasted defaulters are false positives. The recall approximately 79.8%, shows model accurately recognizes approximately 80% of real defaulters thus 20% of defaulters are misclassified as non-defaulters (FN = 404). False Negatives (404) means these are actual defaulters which predicted as non-defaulters as "risk to the bank". False Positives (201) means these are non-defaulters forecasted as defaulters, possible lost business or disallowed loan applications.

10. Discussion

The results of the current study are summarized in the Table 3. The Table 3 provided model performances with values of Accuracy, F1-score, ROC AUC, Precision and Recall. The sum-ups of the model evidently display performance of each model on the basis of testing dataset.

Table 3: Performance of various ML models for loan default prediction.

Model	Accuracy	F1-Score	ROC AUC	Precision	Recall
Random Forest	0.930	0.830	0.975	0.896	0.773
Gradient Boosting	0.926	0.823	0.973	0.885	0.768
XGBoost	0.932	0.837	0.979	0.890	0.789
LightGBM	0.933	0.841	0.979	0.888	0.798

Accuracy: Considered four models achieved admirable accuracy, which displays vigorous whole predictive competence. The model LightGBM has been noted with highest accuracy value as 0.933, trailed meticulously by model XGBoost (0.932), model Random Forest (0.930), and model Gradient Boosting (0.926). These protests the ensemble tree-based representations are enormously capable for loan default forecasting.

F1-Score: The F1-Score equipoises precision and recall, mostly important for excessive datasets. At this time the model LightGBM attained top outcome with (0.841), displays that it is capable at correctly identifying default and non-default properties. The model XGBoost (0.837) and model Random Forest (0.830) the same attained thorough, by model Gradient Boosting slightly minor at 0.823.

ROC AUC: the whole models achieved good ROC AUC marks approximately 0.97, suggesting vigorous justification among defaulters and non-defaulters. The model XGBoost and LightGBM verified highest (0.979), depicting these models are premium at situation of debtors by defaulting risk.

Precision and Recall: the precision assess correctly forecast defaults out of whole forecast defaults; however, recall assesses correctly forecast defaults out of actual defaults. The model LightGBM achieved the highest recall (0.798), as it identifies as extensively held of actual defaulters, while model Random Forest had the highest precision (0.896), observing few false positives. The model XGBoost provides a steady performance by precision value (0.890) and recall value (0.789).

10. Conclusions

From the performance metrics, it is observed that that entire models' performances are outstandingly on the utilised dataset. Even the consequence of the individual models is very good, thus hypothesis for the study is accepted. It is noted that LightGBM model somewhat outperformed other models in almost each metrics. The model LightGBM displays its capabilities predominantly in the handling of class imbalance vis high F1-Score and recall. The model XGBoost is completely follows LightGBM, while model Random Forest surpasses in precision. The consequences of the model performance provide an outstanding inside that advanced ensemble algorithms are very good operative in predicting loan defaults, permitting monetarist units to identify high-risk debtors exactly. Therefore, the supposition of the study is acknowledged with declaration that the machine learning models are meaningly led the classical methods to forecast the loan defaults. As per the confusion matrix approach, it is clear that all above discussed results are based on the confusion matrix. It is one of the base tools to evaluate performance of the models, various evaluation matrix elements are derived from the same. Thus, the accuracy of confusion matrix and its interpretation in the machine learning is too much crucial, as inaccurate confusion matrix might distort the bigger section of assessment.

Conflict of Interest

There is no conflict of interest.

Supporting Information

Not applicable

Use of artificial intelligence (AI)-assisted technology for manuscript preparation

The authors confirm that there was no use of artificial intelligence (AI)-assisted technology for assisting in the writing or editing of the manuscript and no images were manipulated using AI.

References

[1] V. Ivashina, D. Scharfstein, Bank lending during the financial crisis of 2008, *Journal of Financial Economics,* 2010, **97**, 319–338, doi: 10.1016/j.jfineco.2009.12.001.

[2] L. Cheng, T. Kwabena Nsiah, O. Charles, A. Lincoln Ayisi, Credit risk, operational risk, liquidity risk on profitability, A study on South Africa commercial banks, *A PLS-SEM Analysis*, doi: 10.24205/03276716.2020.1002.

[3] N. Uddin, M. K. Ahamed, M. Ashraf Uddin, M. Manwarul Islam, M. Alamin Talukder, S. Aryal, An ensemble machine learning based bank loan approval predictions system with a smart application, *International Journal of Cognitive Computing in Engineering*, 2023, **4**, 327-339, doi: 10.1016/j.ijcce.2023.09.001.

[4] X. Zhu, Q. Chu, X. Song, P. Hu, L. Peng, Explainable prediction of loan default based on machine learning models, *Data Science and Management*, 2023, 6, 123-133, doi: 10.1016/j.dsm.2023.04.003.

[5] X. Zhang, T. Zhang, L. Hou, X. Liu, Z. Guo, Y. Tian, Y. Liu, Data-driven loan default prediction: a machine learning approach for enhancing business process management, *Systems*, 2025, **13**, 581, doi: 10.3390/systems13070581

[6] S. Gour, A. Kumar, R. Shandilya, D. Sharma, Algorithmic approaches for data mining & machine learning, *International Journal of Distributed Computing and Technology*, 2020, 6, 24-29.

[7] W. Wu, Machine learning approaches to predict loan default, *Intelligent Information Management*, 2022, **14**, 157-164, doi: 10.4236/iim.2022.145011.

[8] Y. Sabla, S. Gour, Social media networking analytics and growth perspectives. In: Joshi, A., Mahmud, M., Ragel, R.G., Karthik, S. (eds) ICT: Innovation and Computing. ICTCS 2023. Lecture Notes in Networks and Systems, vol 879. Springer, Singapore, doi: 10.1007/978-981-99-9486-1_22.

[9] K. Bhutani, S. Gaur, P. Panwar, S. Garg, A Neutrosophic Cognitive Maps Approach for Pestle Analysis in Food Industry. In: Rathore, V.S., Tavares, J.M.R.S., Piuri, V., Surendiran, B. (eds) Emerging Trends in Expert Applications and Security. ICE-TEAS 2023. Lecture Notes in Networks and Systems, 2023, 681. Springer, Singapore, doi: 10.1007/978-981-99-1909-3_30.

[10] S. Gour, A Perspective of Industry 4.0, *International Journal of Engineering and Designing Innovation*, 2020, **2**, 1-4.

[11] A. Kumar, S. Gaur, R. Shandilya, D. Sharma, Recommendation system: an algebraic perspective of machine learning with knowledge endorsement, *International Journal of Distributed Computing and Technology*, 2020, **6**, 30-34, 2020, doi: 10.37628.

[12] A. Rajput, S. Agrawal, T. Dua, S. Gour, Plant leaf diseases prediction using convolutional neural network (CNN). In: Rathore, V.S., Manuel R. S. Tavares, J., Tuba, E., Devedzic, V. (eds) Emerging Trends in Expert Applications and Security. ICE-TEAS 2024. Lecture Notes in Networks and Systems, 2024, 1030. Springer, Singapore, doi: 10.1007/978-981-97-3745-1_10.

[13] H. Chen, Prediction and analysis of financial default loan behaviour based on machine learning model, *Hindawi-Computational Intelligence and Neuroscience*, 2022, **2022**, 7907210, doi: 10.1155/2022/7907210

[14] A. Kinjole, O. Shobayo, J. Popoola, O. Okoyeigbo, B. Ogunleye, Ensemble-based machine learning algorithm for loan default risk prediction, *Mathematics*, 2024, **12**, 3423, doi: 10.3390/ math1221342.

[15] L. Zhua, D. Qiua, D. Ergua, C. Yinga, K. Liub, A study on predicting loan default based on the random forest algorithm, Procedia Computer Science, 2019, **162**, 503–513, doi: 10.1016/j.procs.2019.12.017.

[16] L. Monje, R. A. Carrasco, M. S. Montañés, Machine learning XAI for early loan default prediction, *Computational Economics*, 2025, doi: 10.1007/s10614-025-10962-9.

[17] L. Barbaglia, S. Manzan, E. Tosetti, Forecasting loan default in europe with machine learning, *Journal of Financial Econometrics*, 2023, **21**, 569–596, doi: 10.1093/jjfinec/nbab010.

[18] M. A. SharafEldin, A. M. Idrees, S. Ouf, A proposed framework for loan default prediction using machine learning techniques, *International Journal of Advanced Computer Science and Applications*, 2025, **16**, 412-425, 10.14569/IJACSA.2025.0160640.

[19] X. Zhang, T. Zhang, L. Hou, X. Liu, Z. Guo, Y. Tian, Y. Liu, Data-driven loan default prediction: a machine learning approach for enhancing business process management, *Systems* 2025, **13**, 581, doi: 10.3390/ systems13070581

[20] Z. Z. Kang, T. S. Yin, S. Y. G. Tan, W. Chien Ng, Loan default prediction using machine learning algorithms, *Journal of Informatics and Web Engineering*, 2025, **4**, 232-244, doi:10.33093/jiwe.2025.4.3.14.

[21] https: //www.kaggle.com/datasets/taweilo/loan approval-classification-data,

[22] P. S. Saini, A. Bhatnagar, L. Rani, Loan approval prediction using machine learning: A comparative analysis of classification algorithms, 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), May 2 023, 1821-1826.

[23] D. Pandya, A. Jadeja, S. Gour, S. B. Trivedi, H. H. Patel, P. U. Jadeja, An Analytical Perspective of Missing Values in Machine Learning. In: Rathore, V.S., Piuri, V., Babo, R., Tiwari, V. (eds) Emerging Trends in Expert Applications and Security. ICE-TEAS 2024. Lecture Notes in Networks and Systems, 2024, 1037. Springer, Singapore, doi: 10.1007/978-981-97-3991-2_24.

[24] S. Sharma, S Gaur, Contiguous agile approach to manage odd size missing block in data mining, *International Journal of Advanced Research in Computer Science*, 2013, **4**, 214-217, doi: 10.26483/ijarcs.v4i11.1954.

[25] J. C. Alejandrino, J. Jr. P. Bolacoy, J. V. B. Murcia, Supervised and unsupervised data mining approaches in loan default prediction, *International Journal of Electrical and Computer Engineering*, 2023, 13, 1837-1847, doi: 10.11591/ijece.v13i2.pp1837-1847.

Publisher Note: The views, statements, and data in all publications solely belong to the authors and contributors. G R Scholastic is not responsible for any injury resulting from the ideas, methods, or products mentioned. G R Scholastic remains neutral regarding jurisdictional claims in published maps and institutional affiliations.

Open Access