International Journal of Engineering and Designing Innovation (IJEDI)
Monthly, Peer-Reviewed, Refereed, Indexed Journal
Volume - 7, Issue – 11 , November - 2025

ISSN(O): 2582-0788
Impact Factor : 7.14

DOIs:10.2019/IJEDI/202511001 --:-- Research Paper / Article / Review

Assessment of Efficiency of Machine Learning Algorithms in Loan-Default Prediction

[1] Sanjay Gour, [2] Vaibhav Khanna
[1, 2] Department of Computer Science,
Maharshi Dayanand Saraswati University, Ajmer Rajasthan
Email - [1]sanjay.since@gmail.com, [2]isaacajmer@gmail.com

Abstract: *Nowadays the loan default prediction is one of the crucial jobs for the institutions which are financially associated. This activity is unswervingly inducing risk management, approval of loan decisions, along with the profitability. Although there are lot of parameters by which traditionally financial institutions are predicting the loan defaults, but these approaches are not enough to handle properly. Also, there is strong need to not to approve the loan to the person who will be loan defaulter. This study focused on the effectiveness of the various machine learning algorithms models including Random Forest, Gradient Boosting, XGBoost and LightGBM. To solve the problem an organized machine learning approach is established, together with data preprocessing, feature engineering, class imbalance handling, model training and evaluation. The implemented models are evaluated by utilising the matrix of accuracy, F1-score, ROC-AUC, Precision and Recall. This paper is an attempt to synchronised with dataset by the appropriate machine learning algorithms to predict the loan default in the efficient manner by using various evaluation matrix.*

Keywords: *Machine Learning, Loan Default, Random Forest, XGBoost, LightGBM.*

1. INTRODUCTION

The speedy progression of data analytics has transformed the financial industry, predominantly in credit risk calculation and loan default prediction. In the classical banking schemes loan default assessment trusted profoundly on statistical methods like logistic regression, credit scoring, and proficient decision. Though, such approaches frequently flop to capture the composite, nonlinear associations inside huge and varied financial datasets. The machine learning is a subdomain of artificial intelligence, has arose as a commanding substitute proficient of analysing huge volumes of all kind of data to make precise and timely predictions about debtors' possibility of defaulting on loans. As it is well known that loan default prediction is a critical procedure for financial organizations, as it directly impacts loaning verdicts and risk moderation tactics.

A loan default happens when a debtor be unsuccessful in planned payments in the decided timeline, which consequential in monetary suffer for the creditor. The capability to forecast such defaults in early payment empowers financial institutions to accomplish risks efficiently by set the enhance lending strategies, and conform with regulatory outlines. Therefore, evolving and assessing the effectiveness of machine learning algorithms / models in this area has developed a main extent of research and real-world application. Evaluating the efficiency of ML models includes defining how well these models might differentiate amid possible defaulters and reliable debtors. The assessment typically comprises quantifiable system of measurement like accuracy, precision, recall, F1-score, along with area under the receiver operating Characteristic. There is need to assess by the multiple measures relatively because as example a model with high accuracy and low recall might flop to recognize numerous real defaulters. So, several performance metrics are cast-off to confirm a stable assessment of predictive competence.

Numerous machine learning algorithms have remained functional to loan default forecast, together with Random Forests, Gradient Boosting Machines, XGBoost and LightGBM. Amid these, ensemble models have exposed greater performance due to their capability to capture multifaceted interactions amid financial features such as credit past, income, stability, income and loan ratio and similar economic features.

2. REVIEW OF LITERATURE

According to the studies from the year 2019 to 2025, collaborative machine learning models particularly XGBoost, LightGBM, Gradient Boosting, and Random Forest reliably outstrip traditional statistical methods for loan default prediction. Some of the review regarding the same are as follows:

Kinjole et al. (2024) used LendingClub loan data with models like Random Forest, SVM, XGBoost, and ADABoost, applying SMOTE variants to balance data. SMOTE+ENNs with XGBoost achieved 90.49% accuracy, while ensemble stacking raised it to 93.7%, showing that balanced data and ensembles improve prediction.

Leticia Monje et al. (2025) applied XGBoost with a surrogate and fuzzy linguistic model on P2P loans (2007–2020), achieving high accuracy with added interpretability for regulators and early default detection.

Zhi Zheng Kang et al. (2025) used Kaggle loan data (148k records) with Random Forest, XGBoost, and LightGBM, using SMOTE for imbalance. LightGBM performed best (Acc 0.9764, Prec 0.9747, Rec 0.9503); key features were interest rate and credit type.

Luca Barbaglia et al. (2023) analyzed 12M European mortgages, finding XGBoost outperformed logistic regression. Interest rate, LTV, and local economy were major predictors, highlighting regional risk variations.

Zhang X. et al. (2025) tested XGBoost, Gradient Boosting, and LightGBM on institutional loan data. Gradient Boosting achieved best accuracy (0.8887), while XGBoost had highest ROC-AUC (0.9714); introduced cost-sensitive threshold tuning for regulation.

Mona Aly SharafEldin et al. (2025) used Egyptian bank loans with Decision Tree, Random Forest, and Gradient Boosting. Decision Tree (Acc 88%) performed best; key predictors were balance, due amount, and delinquency. Feature selection improved interpretability.

Herui Chen (2022) proposed weighted logistic regression with L2 penalty and TF-IDF features on Chinese credit data, improving imbalance handling and accuracy while reducing overfitting.

Lin Zhua et al. (2019) analyzed LendingClub data using Random Forest, SVM, and Logistic Regression with SMOTE. Random Forest performed best, and SMOTE improved class balance and model reliability.

3. OBJECTIVES:

The objective of the study comprises two key direction which are:

- To device appropriate machine learning approaches including Random Forest, Gradient boosting, XG Boost and LightGBM) for forecasting loan defaults.
- To assess the performance of proposed algorithms by utilising suitable assessment metrics like as accuracy, precision, recall, F1-score, and ROC AUC.

3.1 HYPOTHESIS

The hypothesis for the study framed to compare the significance of machine learning algorithms over the traditional methods so it can be framed as: “The machine learning algorithms are significantly led the traditional methods to predict the loan defaults.”

4. METHODOLOGY:

The methodology of the paper comprises five key sections. At the very first stage selection of the dataset is committed in proper manner with proper credit and demographic history of the persons. The second phase is about Data Pre-processing in this section the key consideration is put on handling of missing data, encoding of categorical data by factorization and normalise the numerical data by feature normalization. After these the dataset becomes ready to work with the machine learning algorithms. At the next stage after correlation analysis and assessments of tree analysis training of the machine process will be done with 80:20 ratio. Here 80% dataset is utilised to train the machine and 20% of the dataset remain for the testing.

After train the machine various machine learning algorithm / models are utilised to process the data for evaluating matrix. Here in this study, we are using Random Forest, Gradient Boosting Machine, XGBoost and LightGBM algorithms for the evaluation of predict the loan defaults. At last, the evaluation matrix assessment is done by using matrix accuracy, precision, recall, F1-score, and ROC AUC.

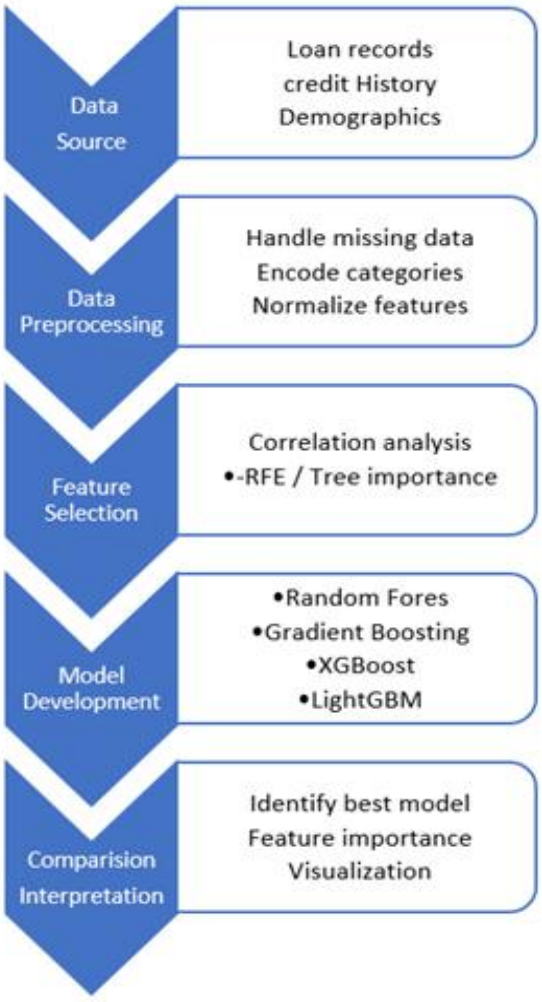

Figure 1: Research Methodology

DATASET

The dataset consider for the study is taken from Kaggle 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.

TOOLS AND TECHNOLOGIES (R STUDIO)

The whole study is accomplished with R programming language; R is an open-source programming platform which is chiefly utilised for data analysis and statistical computing. As per structure point of view it is design to handle data input, processing, and visualization competently. The construction might be mainly alienated into three key components: the R Kernel, R Environment, and R Packages.

The software package RStudio is an integrated development environment (IDE) which is precisely considered for the R language, widely utilised for statistics, data analysis, and machine learning. It delivers a user-friendly interface that streamlines coding, reporting and visualization, projecting it as the platform for both learners and practiced data scientists. The library package as per M L algorithms in R studio are Random Forest “randomForest” and “ranger”, for Gradient Boosting “gbm” and “caret”, for XGBoost library is “xgboost” and for LightGBM library is “lightgbm”.

5. RESULTS:

The result of the study is depicted in the form of model performance which is shown in the table 1. The table gives a consequence of the each and every models. The summaries of the model clearly show high performance of every model on the testing dataset.

Accuracy: the entire models attained excellent accuracy, which representing robust complete predictive capability. The LightGBM model led somewhat with 0.933, trailed thoroughly by XGBoost (0.932), Random Forest (0.930), and Gradient Boosting (0.926). These demonstrations the ensemble tree-based models are extremely competent for loan default prediction.

F1-Score: The F1-Score equilibriums precision and recall, mainly significant for imbalanced datasets. Here the LightGBM another time achieved best (0.841), representing it is competent at properly recognizing default and non-default belongings. The XGBoost (0.837) and Random Forest (0.830) likewise achieved sound, by Gradient Boosting somewhat lesser at 0.823.

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	0.89	0.789
LightGBM	0.933	0.841	0.979	0.888	0.798

Table 1: Performance of ML model for loan default prediction

ROC AUC: the entire models attained outstanding ROC AUC notches upstairs 0.97, signifying robust judicial authority amid defaulters and non-defaulters. The XGBoost and LightGBM recorded uppermost (0.979), representing these models are finest at position borrowers by default risk.

Precision and Recall: the Precision evaluation properly forecast defaults out of total forecast defaults, though recall measures properly forecast defaults out of real defaults. The LightGBM attained the uppermost recall (0.798), representing it recognizes the widely held of real defaulters, whereas Random Forest had the uppermost precision (0.896), viewing some false positives. The XGBoost gives a stable performance through precision (0.890) and recall (0.789).

6. CONCLUSION

It is found that all the model are perform excellently on the given dataset. The result of all the models is individually very good. After assessing and comparing the result from the study it is found that the model LightGBM slightly outperformed other models in furthermost metrics. The LightGBM shows its competences predominantly in the treatment of class imbalance vis higher F1-Score and recall. The XGBoost model is thoroughly follows, whereas Random Forest outshines in precision. The results of the model performance give an excellent inside that progressive ensemble algorithms are extremely effective in predicting loan defaults, allowing financial industries to recognize high-risk borrowers precisely. Thus, the hypothesis of the study is accepted with statement that the machine learning algorithms are significantly led the traditional methods to predict the loan defaults.

REFERENCES :

1. Zhang,X., Zhang,T. ,Hou, L., Liu, X., Guo, Z., Tian, Y. and Liu, Y. (2025), Data-Driven Loan Default Prediction: A Machine Learning Approach for Enhancing Business Process Management, MDPI Systems 2025, vol-13, no-581. https://doi.org/10.3390/ systems13070581
2. Lin Zhua, Dafeng Qiua , Daji Ergua, Cai Yinga, and Kuiyi Liub (2019), A study on predicting loan default based on the random forest algorithm, ELSEVIER-Procedia Computer Science, Science Direct, vol- 162 (2019), Pp-503–513.
3. Zhi Zheng Kang , Teh Sin Yin, Samuel Yong Guang Tan and Wei Chien Ng (2025), Loan Default Prediction Using Machine Learning Algorithms, Journal of Informatics and Web Engineering, vol- 4, no-3, Pp-232-244, DOI:10.33093/jiwe.2025.4.3.14

4. Luca Barbaglia, Sebastiano Manzan and Elisa Tosetti (2023), Forecasting Loan Default in Europe with Machine Learning, OXFORD-Journal of Financial Econometrics, 2023, Vol.-21, No.-2, Pp-569–596. https://doi.org/10.1093/jjfinec/nbab010
5. Mona Aly SharafEldin1, Amira M. Idrees and Shimaa Ouf (2025), A Proposed Framework for Loan Default Prediction Using Machine Learning Techniques, International Journal of Advanced Computer Science and Applications, vol.-16, no. 6, Pp-412-425, 2025
6. Herui Chen (2022), Prediction and Analysis of Financial Default Loan Behaviour Based on Machine Learning Model, Hindawi- Computational Intelligence and Neuroscience, vol-2022, Article ID 7907210, Pp-1-10, https://doi.org/10.1155/2022/7907210
7. Leticia Monje1, Ramón Alberto Carrasco and Manuel Sánchez Montañés (2025), Machine Learning XAI for Early Loan Default Prediction, Computational Economics, Springer, https://doi.org/10.1007/s10614-025-10962-9
8. kinjole, A., Shobayo, O., Popoola, J., Okoyeigbo, O. and Ogunleye, B. (2024) Ensemble-Based Machine Learning Algorithm for Loan Default Risk Prediction, MDPI- Mathematics 2024, vol-12, 3423. https://doi.org/10.3390/ math1221342
9. P. S. Saini, A. Bhatnagar, and L. Rani (2023), Loan approval prediction using machine learning: A comparative analysis of classification algorithms, in 2023 3rd Int. Conf. Adv. Comput. Innov. Technol. Eng. (ICACITE), May 2 023, pp. 1821-1826.
10. Jovanne C. Alejandrino1, Jovito Jr. P. Bolacoy1 and John Vianne B. Murcia (2023), Supervised and unsupervised data mining approaches in loan default prediction, International Journal of Electrical and Computer Engineering (IJECE), April 2023, Vol.-13, Pp.-1837-1847.
11. https: //www.kaggle.com/datasets/taweilo/loan approval-classification-data,