International Journal of Emerging Research in Engineering and Technology
Pearl Blue Research Group| Volume 6 Issue 4 PP 18-26, 2025
ISSN: 3050-922X | https://doi.org/10.63282/3050-922X.IJERET-V6I4P103

Original Article

Data-Driven Loan Default Prediction: Enhancing Business Process Workflows with Machine Learning

Pooja Chandrashekar
Independent Researcher.

Received On: 16/08/2025 ***Revised On: 28/08/2025*** ***Accepted On: 21/09/2025*** ***Published On: 09/10/2025***

Abstract *- Financial institutions perform a major activity, loan default prediction, which directly affects risk management, loan approval and profitability. The study presents a powerful method of loan default forecasting on the Lending Club data, comprising 2,260,701 individual loans records between 2007 and 2018. The presented methodology comprises data preprocessing, such as cleaning, missing values, numerical values scaling, one-hot encoding categorical variables, and correlation and importance-based feature selection. Synthetic Minority Oversampling Technique (SMOTE) is implemented to ensure that the instances of default and non-default are represented in a manner that is class-balanced, thereby mitigating the explicit class imbalance. With this data split into training and test sets, and run ensemble machine learning classifiers, including XGBoost and Light Gradient Boosting Machine (LGBM), to make predictions. Accuracy (ACC), precision (PRE), recall (REC), F1-score (F1), and ROC-AUC scores are the metrics utilized to assess performance. It has been experimentally demonstrated that the developed models possess the excellent performance of prediction, with LGBM reaching a record of 99.99% on all measures and XGBoost coming close with 99.96%- 99.98%. The improved efficacy and generalization capability of the ensemble-based developed approach is proved by comparison with more standard models, including Neural Networks, Support Vector Machines, and Convolutional Neural Networks. The findings indicate that both LGBM and XGBoost offer a highly reliable and interpretable solution to assess financial risk that can identify the possible loan defaults with the best level of accuracy and can contribute to the process of lending choices.*

Keywords *- Loan Default Prediction, Financial, Business, Machine Learning, Banking.*

1. Introduction

Banks all over the world must safeguard their loans books and minimize defaults in order to have healthy financials and sustainability over the long term [1]. In order to achieve this objective, they employ various risk assessment techniques that span both ancient credit rating approaches to complex strategies that integrate AI and ML. The correct segmentation of customers on this basis and based on credit worthiness or the capacity to repay the loans is therefore a significant process before loans are granted. Financial intermediaries that take deposits and use those funds for lending are known as banks. The recent financial landscape has seen predictive analytics becoming an essential instrument of maximizing loan portfolio management [2][3]. Banks, having so much data in form of loan applications, credit histories, and macroeconomic data available, are in a better position to use predictive models to estimate credit risk, forecast potential default, and lend in a way suited to their objectives. This data-driven approach allows institutions to make better decisions, which in turn lowers their risk of financial losses and boosts their profitability. Also, organizations that use predictive analytics can adapt themselves proactively to the changes in the market ensuring profitable financial results.

Despite the intensive expansion of the lending business into a major source of income among the financial institutions, the challenge of the verifiability of the borrowers remains [4][5]. The conventional credit rating and assessment procedures often do not identify high-risk applicants in the right way. Therefore, there is the need to develop a more robust and precise model of risk assessment in prediction of loan defaults [6]. Since lending is a basic banking product whose revenue is mainly generated through the interest rates, the ACC in default prediction directly affects the financial growth and efficiency in operations. Profitability, risk management, and sound decision-making all benefit from accurate loan default prediction. The impact on the balance sheet of a bank by loan default occurs when the borrower is unable to fulfil his or her loan repayment obligations [7][8]. Even though there are extensive datasets of customer demographics, financial behaviour and loan information, their accurate prediction is difficult due to imbalance in the data, heterogeneity in the behaviours and interdependence of multiple financial behaviour. The machine learning techniques have come to the limelight as powerful instruments of financial predictive modelling. ML algorithms can improve prediction ACC because, in contrast to conventional statistical methods, they can identify complicated and non-linear patterns in data.

The increasing number and size of financial transactions, has posed an increased threat of loan defaults, to which accurate forecasting is a critical requirement of lending institutions [9]. The classical statistical and machine learning methods cannot identify the complicated relationships in the high-dimensional financial data, and have the problem of class imbalance that makes their prediction performance poor. The need is great to have powerful, scalable, and interpretable prediction models to be able to handle large-sized datasets with a high degree of efficiency, risk evaluation, and making well-informed decisions within the lending procedure. The motivation behind the same is to develop a state-of-the-art machine learning framework that achieve better predictive ACC, data imbalance, and provide practical insights into financial risk management. The chief contributions of this work are as follows:

- A large-scale loan data specific end-to-end pre-processing pipeline including cleaning, feature selection, scaling and one-hot encoding.
- Effective control of strong class imbalance through the SMOTE to boost model performance on minority default cases.
- Application of ensemble ML models, LGBM and XGBoost, for precise and explainable loan default prediction.
- A versatile and effective infrastructure can greatly assist financial organizations in improving their decision-making and credit risk evaluation procedures.
- Using standard criteria for evaluation (ACC, PRE, REC, F1), the study verifies the efficacy of the models.

The increasing rate of loan default presents vast financial risks to the lenders and, therefore, accurate and reliable predictive models are urgently required. Traditional methods, such as NN, SVM, and CNN, do not typically work with imbalanced data and high-dimensional financial characteristics, and these methods perform poorly. These problems are addressed in this paper by integrating effective data pre-processing, feature selection and balancing of the classes with the use of ensemble based ML algorithms. The innovation is in the use of LGBM and XGBoost on an expanded Lending Club data set with near-perfect prediction scores while retaining model interpretability and generalization. With the integration of strong ensemble learning and rigorous management of data unbalance and feature importance, the proposed method greatly surpasses traditional models in terms of effectiveness, offering a highly efficient and scalable method of accurate loan default prediction and improved financial risk management.

1.1. Structure of Paper

The following is the paper's outline Following a brief literature review in Section II, the methodology is described in Section III. Results and evaluation are presented in Section IV. Finally, suggestions for future study are offered in Section V.

2. Literature Review

This section discusses studies on predicting loan defaults to improve company using ML methods. Table I shows a summary of these studies. Shah (2025) uses machine learning methods for binary classification and regression analysis to assess financial risk factors that determine loan approval dataset. According to the experimental data, LightGBM obtains the highest classification ACC of 96.23% while maintaining a good balance between PRE and REC. In the regression tests, CatBoost performs better, recording the lowest prediction errors as well as the highest R2 score of 0.8820 [10]. Istia et al. (2025) accurately and effectively predict loan approval outcomes. The dataset utilized for this research aims to forecast the approval or rejection of a loan request based on multiple variables. RF model achieved 94.10% of F1, ACC of 94.12%, PRE and REC of 94.24% and 94.12%, respectively, after applying SMOTE for data equilibrium maintenance [11].

C et al. (2024) Loans are playing an increasingly dominant role in the banking industry, but conventional techniques of evaluating them on the basis of mostly asset value and income do not adequately determine good loan borrowers. The study bridges this shortcoming by employing machine learning methods to better predict loan eligibility. Utilize algorithms like Decision Tree, Extra Trees, XGBoost, and LightGBM, with LightGBM reflecting an optimal ACC of 98.91%, in order to handle data and generate more reliable predictions [12]. Kumar Jain et al. (2024) offers a predictive analytics solution that uses open P2P loan data from Lending Club and is based on ML to improve the ACC of loan default prediction in the banking industry. After evaluating LR, DT, and SVM, the study eventually found that RF was the most effective ML model, with 89% ACC, 99.5% sensitivity, and 80.3% specificity which are excellent measures [13].

Chauhan (2024) was conducted in order to solve this challenge. The main objective of this study is to have a better understanding of how to use sophisticated machine learning classification algorithms to predict when bank loans would default. Attained ACC of 88%, 89%, and 87%, respectively, using a variety of classifiers such decision trees, logistic regression, and random forests [14]. Nancy Deborah et al. (2023) utilize DT and KNN algorithms to predict a consumer's loan approval status. The model can be used with the help of a novel technique called the Support Vector Classifier, which demonstrates improved ACC. The study's primary objective was to develop methods for predicting loan status through the application of machine learning techniques. A test algorithm that achieved an impressive 83% ACC score was SVM [15].

Lakshmanarao et al. (2023) created a system that uses ML and DL models to predict when loans would go into default. For this project, used Lending Club's data on defaulted loans. In order to get a pre-processed dataset, the dataset is subjected to multiple data pre-processing techniques. Afterwards, four ML methods—decision trees, logistic regression, K-NN, and feed forward neural networks—were suggested. Experimental results showed that the proposed feed forward neural network has a very high accuracy rate (99%) in forecasting loan defaults [16].

Table 1: Comparative Analysis for loan default Prediction using machine learning

Authors	Methodology	Data	Key Findings	Limitations	Future Work
Shah (2025)	Machine learning for binary classification and regression with Bayesian Optimization for hyperparameter tweaking utilizing LightGBM, Random Forest, and CatBoost.	Loan approval dataset containing financial risk factors.	LightGBM achieved 96.23% ACC in classification; CatBoost had the best regression performance with R² = 0.8820.	Limited interpretability of ensemble models; dataset domain restricted to financial risks only.	Integrate explainable AI (XAI) techniques and include real-time decision systems for risk assessment.
Istia et al. (2025)	Random Forest model with SMOTE for class imbalance; compared 9 ML models including DL and ensemble methods.	Loan approval dataset (binary classification: approved/denied).	RF achieved 94.12% ACC, 94.10% F1, 94.24% PRE, and 94.12% REC.	Focused mainly on RF; lacks interpretability and feature importance analysis.	Extend study using hybrid DL–ML architectures and cross-validation across different datasets.
C et al. (2024)	ML algorithms (Decision Tree, Extra Trees, XGBoost, LightGBM) applied for loan eligibility prediction.	Banking loan dataset with borrower and financial features.	LightGBM achieved 98.91% ACC, outperforming others.	Absence of feature correlation or explainability analysis.	Explore ensemble stacking and integrate macroeconomic indicators.
Kumar Jain et al. (2024)	Predictive analytics on P2P Lending Club data using Random Forest, LR, DT, and SVM for loan default prediction.	Open P2P loan dataset (Lending Club).	RF achieved 89% ACC, 99.5% sensitivity, 80.3% specificity.	Imbalanced dataset handling not detailed; lacks cross-dataset validation.	Use deep learning for temporal risk prediction and improved data balancing.
Chauhan (2024)	Ensemble ML classifiers (LR, RF, DT) for non-performing loan prediction.	Bank loan dataset (loan default classification).	Achieved 87% (LR), 89% (RF), 88% (DT) ACC.	Limited to classical ML; no deep learning comparison.	Incorporate ensemble boosting (e.g., XGBoost, CatBoost) and real-time risk monitoring.
Nancy Deborah et al. (2023)	KNN, Decision Tree, and SVM applied to bank loan approval prediction system.	Bank loan approval dataset.	SVM achieved 83% ACC, outperforming others.	Low ACC compared to ensemble methods; no hyperparameter tuning.	Apply ensemble learning and hyperparameter optimization to improve results.
Lakshmanarao et al. (2023)	Loan default prediction using DT, RF, LR, KNN, and Feed Forward Neural Network (FNN).	Lending Club loan default dataset.	FNN achieved 99% ACC, outperforming traditional ML models.	Possible overfitting due to high ACC; dataset from single source.	Validate on other financial datasets and use explainable deep learning models.

3. Methodology

The approach used to forecast loan defaults using the Lending Club dataset follows a systematic series of data processing and modelling stages, as shown in Figure 1. Pre-processing and data cleaning ensures that there is no noise, missing values, or inconsistent entries in the final product. Subsequently, redundant and irrelevant variables are removed, and numerical features are scaled while categorical variables are converted using one-hot encoding. Feature selection methods are thereafter used in loan default. To handle class imbalance, SMOTE is utilized to provide balanced representation of both classes. The next step is to split the cleaned dataset in half, creating a training set and a testing set. From the training set, two machine learning models, XGBoost and LightGBM, are proposed and trained. Finally, the models' predictive power in identifying loan default risk is evaluated using popular measures such as ACC, PRE, REC, and F1.

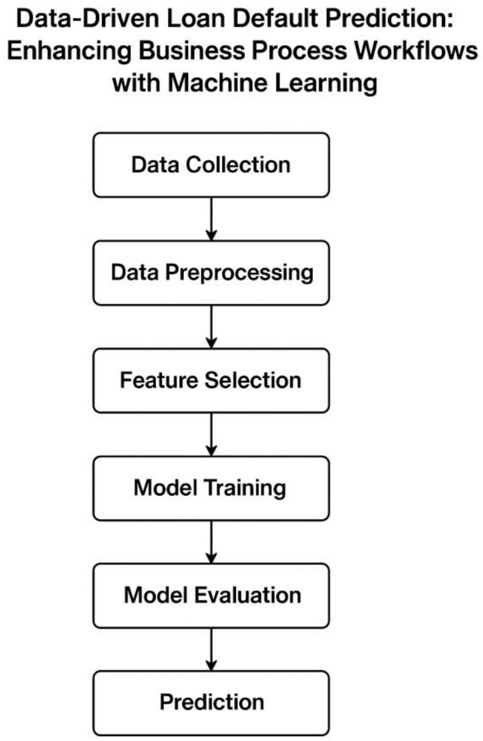

Fig 1: Methodology Flowchart of Loan Default Prediction

The next sections include a full explanation of each stage in the technique and the recommended flowchart:

3.1. Data Collection

This study looks at the numbers behind approved loan applications from 2007 to 2018 using data retrieved from the Lending Club website. There are 151 variables and 2,260,701 observations in the whole dataset. Borrower details, loan information, credit history, and other personal characteristics are the main variables. The dataset was cleansed of all combined applications since this analysis just addresses individual loans. The entire dataset has 2,260,701 observations and 151 variables. The graph for visualization is given below.

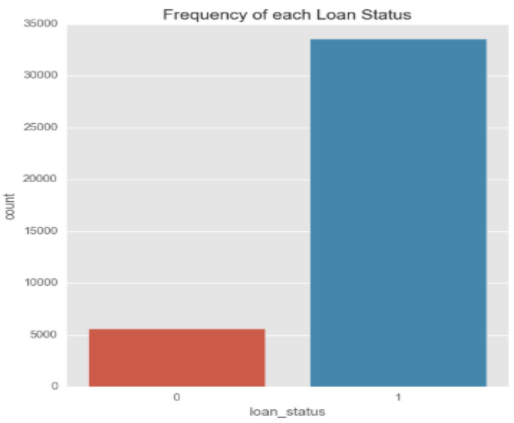

Fig 2: Loan Status Distribution

Figure 2 shows a class imbalance in the distribution of the binary variable loan status, which can take on values of 0 or 1. The class 1 is the dominant one, with a count of almost 33,000 observations, while the class 0 is the minority one, with a count of about 5,500 observations. This significant difference reflects that the dataset is strongly biased toward a single class, which is a critical aspect to balance through the predictive modeling exercise in order to achieve fair model performance.

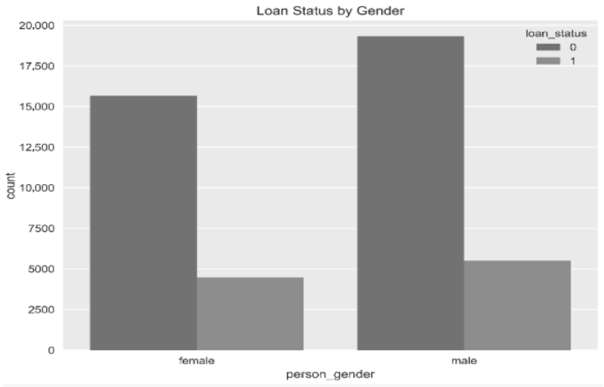

Fig 3: Loan Default Rates by Gender

Figure 3, presents bar chart Loan Status by Gender, is a comparison of the number of loans broken down by gender ('female' and 'male') and loan status. The chart unequivocally presents that loans were given to males (total number close to 20,000) in greater numbers than to females (total number close to 15,000). A considerably larger percentage of loans for both male and female borrowers have a status of 0 (not in default) than a status of 1 (in default), indicating that this is the case for the vast majority of loans.

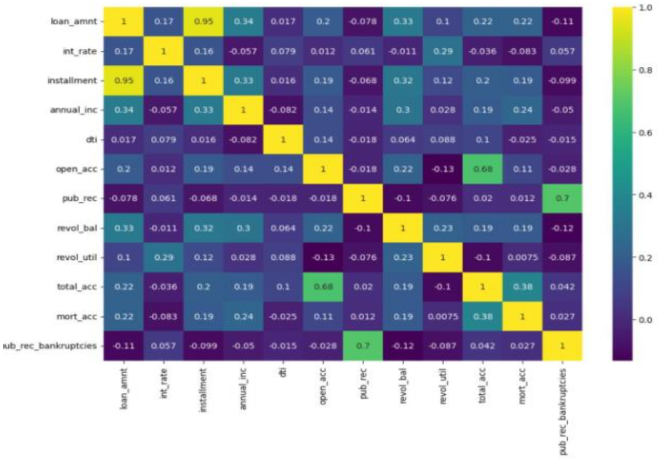

Fig 4: Correlation Heatmap

The Figure 4 is a Correlation Heatmap representing the linear correlations among different financial variables (such as loan_amnt, int_rate, installment, etc.). Strong positive correlation (such as loan_amnt and installment) is represented by bright yellow, while stronger correlations are represented by darker colors. It's utilized for the rapid identification of highly correlated features in the dataset.

3.2. Data Pre-processing

During the data pre-processing phase, multiple actions are taken to transform and prepare the dataset for machine learning. A key step involves converting numerical representations for categorical parameters. The following steps of pre-processing are listed in below:

3.3. Data Cleaning

The raw dataset contains sentences sourced from lending club dataset, which are typically well-formatted. However, basic text cleaning is still necessary. This includes:

- **Removing Redundant and Irrelevant Variables:** A number of redundant variables were eliminated prior to the selection of the final variables for predictive modelling [17]. To begin, extracted all the non-loan-related data, such as the LendingClub membership ID of the borrower.
- **Converting Variables:** The categorical variables were converted to numerical values in order to prepare them for model training. At first, set the dependent variable "Loan status" to have a "non-defaulted" category of 0 and a "defaulted" category of 1.
- **Dealing with Missing Data:** One of the initial steps in processing the training data for the models was to address missing values. In most cases, take an average of the variables and use that to fill in any blanks in the data [18].

3.4. One-Hot Encoding

A frequently employed approach for this objective is one-hot encoding. It entails making binary columns for every possible value of a categorical variable. A value of 1 is assigned to each observation if it falls into a specific category, and a value of 0 is assigned otherwise. Each column represents a different category.

3.5. Feature Selection

The term "feature selection" refers to the steps used to extract the most useful variables from a dataset for use in predictive modelling [19]. By eliminating unnecessary or superfluous characteristics, it aids in dimensionality reduction, boosts model performance, and reduces the likelihood of overfitting. To identify the most important qualities, researchers often use methods like feature importance ranking, correlation analysis, and mutual information.

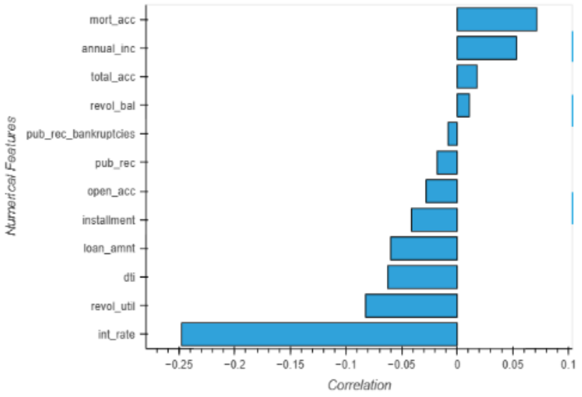

Fig 5: Correlation Between Loan Status

Figure 5 shows the results of transforming the "loan status" variable. The "Defaults" and "Fully Paid" categories were set to 0 and 1, respectively, so that the association with other aspects could be determined. There appears to be a positive correlation between "mort_acc," "total_acc," "annual_inc," and "revol_bal," and a negative correlation between "revol_util," "int_rate," and "loan_amnt." Nonetheless, there isn't a single correlation that shows a really significant relationship.

3.6. Data scaling

The data was standardised to ensure that each variable would only have a proportional impact on the prediction result. This is crucial because the prediction performance can be significantly influenced by the magnitude of the variables when using algorithms that employ the mean square error as the loss function, as models are sensitive to variables with large scales [20].

3.7. Addressing Class Imbalance

The training set was subjected to the SMOTE. To help models learn more accurate representations of minority class distribution, SMOTE creates synthetic samples using existing minority-class observations. This successfully balances the training dataset. For the defaulted loans, this two-pronged strategy greatly improved REC and predicted ACC.

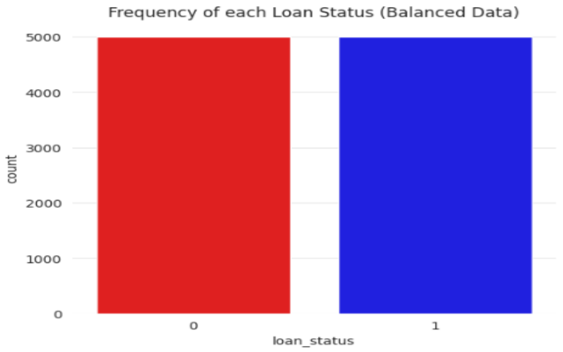

Fig 6: Balancing Graph using SMOTE

In Figure 6, a bar chart titled "Frequency of each Loan Status," show how SMOTE helped level the playing field when it came to the loan status parameter. The chart indicates that both classes, '0' and '1', have almost equal frequencies of about 5,000, reflecting an equilibrium dataset. This equilibrium is a must for training a strong and unbiased classification model that can guarantee accurate prediction of loan defaults.

3.8. Data Splitting

Training comprised 80% of the dataset, while testing accounted for 20%. In order to maintain parity in the dependent variable (loan status), stratified splitting was used.

3.9. Proposed XGBoost and LGBM Models

Machine learning classifiers are algorithms that can be taught patterns and then use that knowledge to generate predictions with the use of labelled training data. The credit scoring, fraud detection, image recognition, and NLP industries are just a few that rely heavily on them. This study made use of the XGB and LGBM ML classifiers. The Light Gradient Boosting Model (LGBM) is one method for gradient boosting that relies on tree-based learning; it is defined in Equation (1). It can handle massive data since it is efficient and scalable [21][22]. Using a gradient-based optimization method, LGBM constructs an ensemble of weak models with the goal of iteratively minimizing the loss function. For better ACC and faster training, it incorporates procedures like leafwise tree growth and histogram-based binning. Coefficients are represented by αi, weak model count is $hi(x)$, and N is the total number of weak models. As seen in Equation (2), XGB (Extreme Gradient Boosting) is another well-known gradient boosting method that offers outstanding predictive ACC. (1). It relies on the same basic premise as LGBM but adds extra regularization strategies to forestall overfitting. Maximizing a differentiable loss function is the goal of XGB, a model that combines gradient boosting with DTs. The objective can be optimized and many evaluation measures can be used with its versatility. The base models are denoted by $fi(x)$, the decision trees are denoted by $T(x;\Theta t)$, and the step sizes are denoted by γt.

$$LGBM: Ensemble(x) = \sum_{i=1}^{N} \alpha_i \times h_i(x) \tag{1}$$

$$XGBoost: Ensemble(x) = \sum_{i=1}^{N} f_i(x) + \gamma_t \times T(x; \Theta_t) \tag{2}$$

3.10. Performance Evaluation Matrix

The models' efficacy was evaluated using a variety of metrics, including the F1, REC, ACC, PRE, and ROC-AUC. In conclusion, the model's prediction was quantified using accuracy. A combination of ACC, PRE, and REC was used to evaluate the model's accuracy in making correct class predictions. The F1, averaging out PRE and REC, was especially useful where there had been an imbalance in class distribution [23]. The application of the ROC-AUC metric allowed us to determine how accurately the model had differentiated between positive and negative classes based on different levels of classification thresholds. Comparative judgments of performance based on these measures, were able to better determine the most appropriate model to apply for loan risk forecasting. The all metrics is obtained in Equation (3-6):

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

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * (precision * recall)}{precision + recall}$$

TP refers to cases when the model correctly anticipates a loan that has defaulted, whereas TN denotes cases where the model correctly identifies a loan that has not defaulted. The term "False Negative" (FN) refers to a loan that has already defaulted and "False Positive" (FP) describes situations where a non-defaulted loan is mistakenly thought to have defaulted.

4. Results and Discussion

An Intel (R) Core (TM) i5 processor with 3.20 GHz, 16 GB of RAM, and Windows 10 Pro are all needed to run the experiments. To put the suggested paradigm into action in Python 3.7.3, the software specification calls for Jupyter Notebook (Anaconda3). The model has been pre-processed using a number of packages, such as sklearn and Pandas. The Lending Club loan dataset was used to assess the performance of the suggested models, XGBoost and LGBM, for the prediction of loan defaults. Table II provides a summary of the findings. The findings show that both models have very high predictive ACC, where LGBM reaches 99.99% in ACC, PRE, REC, and F1, whereas XGBoost is very close with 99.96%, 99.98%, 99.97%, and 99.98%, respectively.

Table 2: Performance of propose Model on lending club loan dataset for loan default

Performance Matrix	LGBM	XGBoost
Accuracy	99.99	99.96
Precision	99.99	99.98
Recall	99.99	99.97
F1-score	99.99	99.98

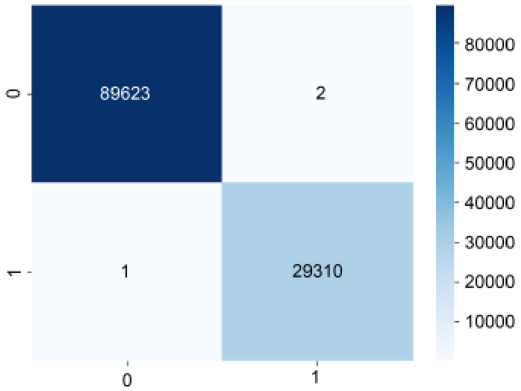

Fig 7: The Confusion Matrix LGBM Model

The LGBM model's confusion matrix for the loan default prediction task is shown in Figure 7. The model has excellent performance in terms of classification with 89,623 TN and 29,310 TP suggesting the large capacity to distinguish between non-defaulters and defaulters. Moreover, the model only registered FP and FN, which indicates few cases of misclassifications and indicates the high accuracy and reliability of the model in predicting the outcomes of loan default.

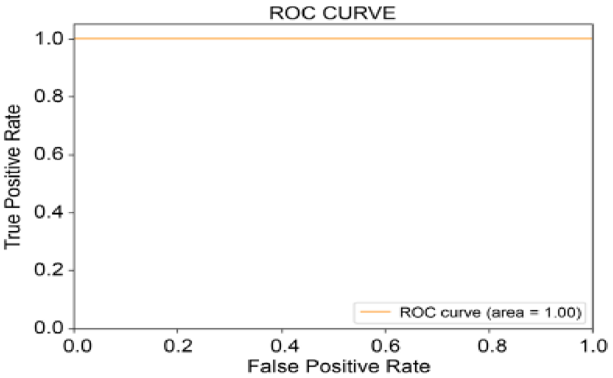

Fig 8: ROC Graph of LGBM Model

The ROC curve shown in Figure. 8 represents the work of the LGBM model on the loan default classification problem. The graph compares the TPR and the FPR in order to make an extensive assessment of the model discriminative capacity. Positioned in the upper-left corner of the plot, the orange ROC curve indicates that the classification performance is around ideal. The gradient of the curve as indicated in the legend is 1.00 and represents an ideal classifier.

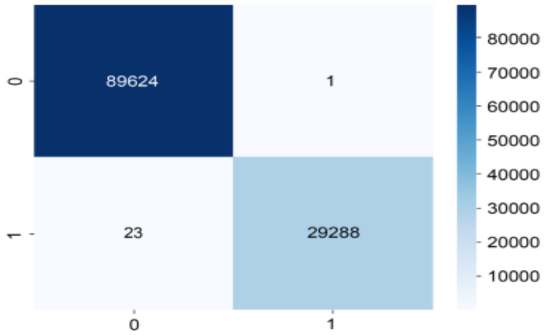

Fig 9: Confusion Matrix of XGBoost

Figure 9 presents the confusion matrix to indicate how the XGBoost model performed on the loan default classification task. The model has a very high predictive ACC with TN and TP correctly forecasting 89,624 and 29,288 instances, respectively, of non-defaulters and defaulters. There were low misclassifications of 1 FP, that is, non-defaulter was a defaulter, and 23 FN, that is, defaulters were non-defaulters.

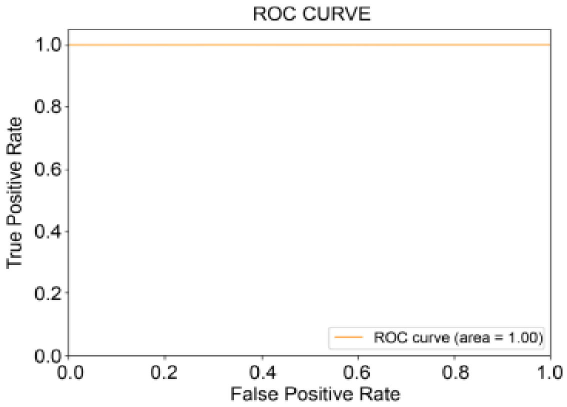

Fig 10: ROC Curve Graph of XGBoost

The ROC curve as presented in Figure. 10 is the performance analysis carried out by the XGBoost model on the task of classifying loan defaults. The curve is used to plot TPR verses FPR to represent the tradeoff between specificity and sensitivity at different threshold values. The orange ROC curve placed on the upper end of the plot with a TPR of 1.0, which is close to ideal classification results. The AUC, as evidenced in the legend, is 1.00, which is an ideal classifier.

Table 3: Comparison between base and propose model Performance matrix for Loan default Prediction

Performance Matrix	LGBM	XG Boost	NN [24]	SVM [25]	CNN [26]	Stacking A [27]
Accuracy	99.99	99.96	91	77.6	67.27	93.69
Precision	99.99	99.98	88	81	68.24	95.59
Recall	99.99	99.97	89	46	64.94	95.55
F1-score	99.99	99.98	-	59	66.43	97.81

Table III demonstrates comparative analysis of the proposed LGBM, with the XGBoost model, with the existing methods, which are the NN, SVM, CNN and Stacking Model A, in predicting loan default. According to the results, the proposed models outperform the baseline methods across the board in terms of performance indicators. LGBM model recorded the highest values with an ACC, PRE, REC and F1 of 99.99%, 99.98%, 99.97% and 99.98% respectively and the XGBoost model has the nearest values with 99.96%, 99.98%, 99.97% and 99.98% respectively. Conversely, the conventional models like NN, SVM and CNN had lower ACC rates of 91%, 77.6% and 67.27% respectively, which showed a high predictive power and strength of the generalization features of the proposed ensemble-based models in predicting loan default risks.

The suggested model utilizes the LGBM and XGBoost to improve the ACC and reliability of loan default forecasting. It effectively uses gradient boosting methods to operate on large and complex data sets as well as minimizing overfitting. Compared with the traditional methods, the model is more accurate, recalls higher and has a higher F1-score, is faster to train, easier to interpret, and has a higher generalization. Its strong output renders it very efficient in the assessment of financial risk and loan decision-making.

5. Conclusion and Future Scope

This paper suggests a powerful and effective architecture of loan default forecasting on the Lending Club dataset (20072018) with the help of the LGBM and Extreme Gradient Boosting (XGBoost) higher-order frameworks of the ensemble-based machine learning. The methodology implies the systematic assembly of data pre-processing, feature selection, class balancing (through SMOTE), and model assessment to guarantee the quality of data as well as fairness and predictive ACC. The results of the experiment confirm that both models provide close to perfect classification with LGBM showing 99.99 of ACC, PRE, REC and F1 and XGBoost showing right behind with 99.96%-99.98 of all the measures. Comparative analysis also indicates that the gradient boosting models have been found to be much better than the conventional ML and DL techniques like NN, SVM, as well as CNN.

5.1. Recommendations

Lending organizations and financial institutions are advised to consider an ensemble-based machine learning framework, such as LGBM and XGBoost, to predict loan defaults in real-time. Introducing such models into credit scoring systems and automated credit pipelines of loan awarding can greatly enhance the degree of risk assessment. Moreover, the implementation of such predictive systems into an existing monitoring infrastructure will enable the institutions to dynamically evaluate the behavior of the borrowers and to adjust risk levels as time goes on. They also suggest that banks should frequently update their models using new financial and behavioural data in order to keep the models relevant and predictively stable.

5.2. Limitations of the Study and Future Work

The proposed LGBM and XGBoost models have remarkable predictive power, however they are not without their limits. The dataset was drawn from 2007 to 2018, and may not fully account for the most recent lending practices, economic changes, and emerging borrower behaviours in the post-COVID-19 period, which might limit the generalizability of the model. The empowering factor in the present study was focusing solely on structured tabular data, thus external factors that can affect the way lending works were not included, such as macroeconomic indicators at the time, social behaviour, and sentiment analysis, which would have helped in producing a more accurate predictive risk score. Nevertheless, high ACC does indicate reasonably strong performance; however, it does raise some concern with overfitting in the models. Next steps for the future research agenda would be to integrate Explainable AI (XAI) methods, such as SHAP, to enhance the interpretability of the data, integrate real-time financial and behavioural data for prediction, integrate temporal or hybrid deep learning models for enhanced adaptation, and utilize cloud-based or API systems to implement models for predictive risk score application. Deploying into a real banking system will allow for continuous learning and a working implementation applied into a multifaceted financial system.

References

[1] I. R. Berrada, F. Barramou, and O. B. Alami, “Towards a Machine Learning-based Model for Corporate Loan Default Prediction,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 3, pp. 565–573, 2024, doi: 10.14569/IJACSA.2024.0150357.

[2] A. Tripathi, “Data-Driven Predictive Analytics For Loan Portfolio Management : Proactive Decision-Making .,” *Int. J. Creat. Res. Thoughts (IJCRT*, vol. 13, no. 5, pp. 916–923, 2025.

[3] S. R. Kurakula, “The Role of AI in Transforming Enterprise Systems Architecture for Financial Services Modernization,” *J. Comput. Sci. Technol. Stud.*, vol. 7, no. 4, pp. 181–186, May 2025, doi: 10.32996/jcsts.2025.7.4.21.

[4] A. Al-Qerem, G. Al-Naymat, M. Alhasan, and M. Al-Debei, “Default prediction model: The significant role of data engineering in the quality of outcomes,” *Int. Arab J. Inf. Technol.*, 2020, doi: 10.34028/iajit/17/4A/8.

[5] G. Modalavalasa and S. P. Bheri, “Next-Generation AI-Powered Automation for Streamlining Business Processes and Improving Operational Efficiency,” *J. Comput. Technol.*, vol. 12, no. 12, pp. 1–7, 2023.

[6] K. B. Thakkar and H. P. Kapadia, “The Roadmap to Digital Transformation in Banking: Advancing Credit Card Fraud Detection with Hybrid Deep Learning Model,” in *2025 2nd International Conference on Trends in Engineering Systems and Technologies (ICTEST)*, 2025, pp. 1–6. doi: 10.1109/ICTEST64710.2025.11042822.

[7] X. Zhang *et al.*, “Data-Driven Loan Default Prediction: A Machine Learning Approach for Enhancing Business Process Management,” *Systems*, vol. 13, no. 7, 2025, doi: 10.3390/systems13070581.

[8] A. R. Bilipelli, “Forecasting the Evolution of Cyber Attacks in FinTech Using Transformer-Based Time Series Models,” *Int. J. Res. Anal. Rev.*, vol. 10, no. 3, pp. 383–389, 2023.

[9] V. Verma, “Deep Learning-Based Fraud Detection in Financial Transactions : A Case Study Using Real-Time Data Streams,” vol. 3, no. 4, pp. 149–157, 2023, doi: 10.56472/25832646/JETA-V3I8P117.

[10] S. B. Shah, “Advanced Framework for Loan Approval Predictions Using Artificial Intelligence-Powered Financial Inclusion Models,” in *2025 IEEE Integrated STEM Education Conference (ISEC)*, 2025, pp. 1–10. doi: 10.1109/ISEC64801.2025.11147327.

[11] U. A. M. Istia, Al-Amain, K. M. M. Uddin, M. T. Ul Islam, and M. A. Based, “An Integrated Approach Using Ensemble Machine Learning and Deep Learning for Loan Approval Prediction,” in *2025 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, IEEE, Feb. 2025, pp. 1–6. doi: 10.1109/ECCE64574.2025.11013889.

[12] S. K. C, M. S, P. M. Reddy, and K. Gopal, “Analyzing the Performance of Ensemble Machine Learning Algorithms for Predicting Loan Eligibility,” in *2024 9th International Conference on Communication and Electronics Systems (ICCES)*, IEEE, Dec. 2024, pp. 1362–1367. doi: 10.1109/ICCES63552.2024.10859945.

[13] Y. K. Jain, P. K. Mannepalli, K. Kaur, A. Maheshwari, and J. Singh, “Effective Machine Learning-Based Predictive Analytics for Loan Default Prediction in Banking Sector,” in *2024 International Conference on Communication, Control, and Intelligent Systems (CCIS)*, IEEE, Dec. 2024, pp. 1–6. doi: 10.1109/CCIS63231.2024.10931843.

[14] S. Chauhan, “Machine Learning Models for Loan Default Forecasting: Accuracy Comparison,” in *2024 Second International Conference Computational and Characterization Techniques in Engineering &amp; Sciences (IC3TES)*, IEEE, Nov. 2024, pp. 1–5. doi: 10.1109/IC3TES62412.2024.10877523.

[15] R. Nancy Deborah, S. Alwyn Rajiv, A. Vinora, C. Manjula Devi, S. Mohammed Arif, and G. S. Mohammed Arif, “An Efficient Loan Approval Status

Prediction Using Machine Learning," in *2023 International Conference on Advanced Computing Technologies and Applications (ICACTA)*, IEEE, Oct. 2023, pp. 1–6. doi: 10.1109/ICACTA58201.2023.10392691.

[16] A. Lakshmanarao, C. Gupta, C. S. Koppireddy, U. V. Ramesh, and D. R. Dev, "Loan Default Prediction Using Machine Learning Techniques and Deep Learning ANN Model," in *2023 Annual International Conference on Emerging Research Areas: International Conference on Intelligent Systems (AICERA/ICIS)*, IEEE, Nov. 2023, pp. 1–5. doi: 10.1109/AICERA/ICIS59538.2023.10420221.

[17] R. Q. Majumder, "Machine Learning for Predictive Analytics : Trends and Future Directions," *Int. J. Innov. Sci. Res. Technol.*, vol. 10, no. 4, 2025.

[18] R. Sifrain, "Predictive Analysis of Default Risk in Peer-to-Peer Lending Platforms: Empirical Evidence from LendingClub," *J. Financ. Risk Manag.*, vol. 12, no. 01, pp. 28–49, 2023, doi: 10.4236/jfrm.2023.121003.

[19] S. J. Wawge, "A Survey on the Identification of Credit Card Fraud Using Machine Learning with Precision, Performance, and Challenges," *Int. J. Innov. Sci. Res. Technol.*, vol. 10, no. 4, May 2025, doi: 10.38124/ijisrt/25apr1813.

[20] G. Mantha, "Transforming the Insurance Industry with Salesforce: Enhancing Customer Engagement and Operational Efficiency," *North Am. J. Eng. Res.*, vol. 5, no. 3, 2024.

[21] A. Aljadani, B. Alharthi, M. A. Farsi, H. M. Balaha, M. Badawy, and M. A. Elhosseini, "Mathematical Modeling and Analysis of Credit Scoring Using the LIME Explainer: A Comprehensive Approach," *Mathematics*, vol. 11, no. 19, p. 4055, Sep. 2023, doi: 10.3390/math11194055.

[22] N. Malali, "Exploring Artificial Intelligence Models for Early Warning Systems with Systemic Risk Analysis in Finance," in *2025 International Conference on Advanced Computing Technologies (ICoACT)*, IEEE, Mar. 2025, pp. 1–6. doi: 10.1109/ICoACT63339.2025.11005357.

[23] H. Kali, "Optimizing Credit Card Fraud Transactions identification and classification in banking industry Using Machine Learning Algorithms," *Int. J. Recent Technol. Sci. Manag.*, vol. 9, no. 11, pp. 85–96, 2024.

[24] N. K. Kokkalakonda, "Risk Assessment In Banking : Ai-Driven Predictive Models For Loan Default Prediction," *Int. Res. J. Mod. Eng. Technol. Sci.*, vol. 7, no. 03, pp. 8623–8633, 2025.

[25] M. A. Kheneifar and B. Amiri, "A Novel Hybrid Model for Loan Default Prediction in Maritime Finance Based on Topological Data Analysis and Machine Learning," *IEEE Access*, vol. 13, no. May, pp. 81474–81493, 2025, doi: 10.1109/ACCESS.2025.3566066.

[26] P. C. Ko, P. C. Lin, H. T. Do, and Y. F. Huang, "P2P Lending Default Prediction Based on AI and Statistical Models," *Entropy*, vol. 24, no. 6, pp. 1–23, 2022, doi: 10.3390/e24060801.

[27] A. Akinjole, O. Shobayo, J. Popoola, O. Okoyeigbo, and B. Ogunleye, "Ensemble-Based Machine Learning Algorithm for Loan Default Risk Prediction," *Mathematics*, vol. 12, no. 21, p. 3423, Oct. 2024, doi: 10.3390/math12213423.