Advancements in Credit Scoring, Profit Scoring, and Portfolio Optimization for P2P Lending

Premkumar Nayaka, Anusha Hegde, Biswajit Bhowmik
Ishwarchandra Vidyasagar AIT Lab, BRICS Laboratory
Department of Computer Science and Engineering
National Institute of Technology Karnataka
Surathkal, Mangalore-575025, Bharat
Email: {sjpremkumarnayaka.232cs030@nitk.edu.in, anushahegde.227cs001, brb}@nitk.edu.in

***Abstract*—Peer-to-peer (P2P) lending sites have revolutionized conventional lending with decentralized, data-driven solutions connecting lenders and borrowers directly. It is essential in such systems to have correct credit risk estimation in order to avoid financial losses and maximize portfolio return. In this paper, a machine learning approach for predicting loan defaults is studied using the Bandura dataset and incorporating borrower, loan, and financial attributes. We run and compare Logistic Regression as the baseline model with Random Forest as an ensemble machine learning method on both binary (default vs. non-default) and multi-class (default, non-default, and prepayment) classification problems. Heavy preprocessing procedures involving feature selection, encoding, normalization, and class handling were utilized in order to promote stable model performance. Our experiments indicate that Random Forest significantly dominates Logistic Regression on all tasks using higher accuracy, precision, recall, and F1 scores. In addition, the native feature importance function of Random Forest introduces interpretability and understanding regarding influential factors on default behaviour. This makes it a dependable and scalable credit risk model for the evolving nature of P2P lending.**

***Index Terms*—Financial Technology, Peer-to-peer Lending, Credit Scoring, Profit Scoring, Portfolio Optimization**

I. Introduction

Peer-to-peer (P2P) lending has radically changed the traditional banking model by providing a decentralized platform alternative to traditional banking. Rather than using banks or financial institutions as middlemen for loan transactions, P2P platforms bring together borrowers and private or institutional lenders directly via online marketplaces. Having direct contact means there is no need for intermediaries, thus saving operational and administrative expenses. Consequently, P2P lending attracts lenders as it can potentially provide higher yields than fixed-income securities or regular savings accounts and, hence, is eligible for portfolio diversification [1]. Borrowers, on the other hand, are rewarded with quicker loan processing procedures, more credit accessibility—particularly to underserved clients of conventional banks—and potentially lower interest rates. Since its inaugural 2000s release, the model has become rapidly popular because of its efficiency, openness, and potential to align the investment objectives of lenders with borrowers' requirements. The sites mostly employ data-oriented algorithms and credit assessment tools for analyzing loan proposals and providing categories of risk, improving decision-making and risk control. In addition, the Internet foundation of P2P lending allows for timely monitoring of loan performance, computerized payment, and efficient user interfaces, which lead to overall convenience [2]. Eliminating middlemen in financial institutions not only facilitates lending but also enables everybody to get financing, thus being a revolutionary part of modern-day financial systems.

Peer-to-peer (P2P) lending has turned out to be a disruptive technology for the financial sector in that it addressed the fundamental financing needs of underbanked populations—like consumers with low credit scores, thin credit reports, or small businesses with no collateral. Traditional financial institutions typically define these borrowers as high-risk and deny them access to credit due to stringent lending criteria. P2P websites fill this vacuum by offering an alternative lending model that connects borrowers with institutional or individual lenders directly without go-betweens and their associated costs. Simple online loan application processes are used by P2P websites, making the borrowing process faster and easier. Loan conditions also usually suit the borrower's financial situation for better accessibility and affordability. For the lenders, P2P platforms offer a new asset class with flexible and diversified investment alternatives to accommodate different risk appetites [3]. P2P platforms give returns higher than regular savings accounts or government bonds, which are coveted by yield-chasing investors. P2P lending is made more successful and appealing by the integration of advanced financial technologies like machine learning algorithms, big data analytics, and credit scoring models. These technologies enable more accurate borrower risk assessment, personalize loan offers, and optimize lender-borrower matching. Therefore, P2P lending has not only made credit accessible to all but also turned out to be a trustworthy, technology-driven alternative to mainstream finance in both consumer and small business segments.

Peer-to-peer (P2P) lending sites use sophisticated data analytics and machine learning (ML) algorithms to improve the accuracy and efficiency of credit assessments. These sites gather and analyze vast amounts of borrower information—ranging from financial history, behaviour patterns, and alternative data sources—to determine creditworthiness and

estimate the likelihood of loan default. By combining such data-driven approaches, P2P platforms can establish fair and risk-adjusted interest rates, which match the risk profiles of borrowers with the investors' return expectations. This methodical and technology-supported process not only provides more precise credit valuations but also enables lenders to make well-informed, evidence-based investment choices [4]. In addition, the automation these technologies provide quickens loan processing procedures, creating quicker turnaround times and a smoother experience for lenders. It also allows for loan product customization, addressing the different needs and financial situations of borrowers. Transparency is another important advantage since platforms generally provide comprehensive and transparent information about loan terms, repayment schedules, and probable risks, and this encourages increased trust and dependability in the lending process. At the core of this process is credit scoring, an integral component that informs borrower evaluation and risk management. New credit scoring algorithms, frequently supplemented by explainable AI methods, play an important role in promoting fairness and accountability in the decision-making process. The synergy of speed, personalization, transparency, and data-driven risk assessment renders P2P lending an attractive alternative to traditional financial systems.

Peer-to-peer (P2P) lending platforms utilize sophisticated data analytics and machine learning (ML) models to improve the accuracy and effectiveness of credit assessments. These platforms gather and analyze large volumes of borrower data—such as financial history, behavioural information, and alternative data sources—to evaluate creditworthiness and forecast the likelihood of loan default. By combining such data-driven approaches, P2P platforms can determine fair and risk-adjusted interest rates, which match the risk profiles of borrowers with investors' return expectations [5]. This systematic and technology-driven method not only provides more accurate credit estimates but also enables lenders to make educated, evidence-based decisions. Unlike other banking systems in which lending risk is undertaken entirely by financial institutions, Peer-to-Peer (P2P) lending platforms share the lending risk with a number of individual investors. This decentralized framework enables investors to directly evaluate prospective borrowers, commonly based on credit scores calculated from a mix of debt-to-income ratios, credit history, and income levels. These scores present a snapshot of a borrower's financial stability and assist in projecting the probability of default. The conventional credit score models, though, have severe shortcomings. These models are generally biased against borrowers who are underbanked or have a limited history of credit repayment, even when they might otherwise be creditworthy [6].

In addition, these models usually over-weight historical payment habits, which cannot necessarily be seen as indicative of a borrower's current financial resources or future reliability. Consequently, the evaluations that result from such models are likely to be inaccurate or poorly calibrated. This, in turn, has fostered an increasing need for more advanced credit risk assessment frameworks that are informed by real-time borrower conduct, alternative data points, and macroeconomic indicators like inflation levels, unemployment rates, or sectoral trends. By combining these dynamic and contextual elements, new risk assessment models can provide a more comprehensive and accurate evaluation of a borrower's creditworthiness, ultimately improving the reliability and inclusivity of P2P lending systems.sed investment choices [1]. In addition, the automation provided by these technologies speeds up loan approval procedures, leading to quicker turnaround times and a more streamlined experience for borrowers.

It also facilitates the customization of loan products, catering to the diverse needs and financial conditions of applicants. Transparency is another key benefit, as platforms typically offer clear and detailed insights into loan terms, repayment plans, and potential risks, fostering greater trust and reliability in the lending environment. To address these limitations, profit scoring has emerged as an innovative framework that prioritizes overall loan profitability over default risk alone. Profit scoring provides a comprehensive view of a loan's financial potential by accounting for interest rates, repayment patterns, and market dynamics. This approach allows lenders to identify high-yield loans that may carry moderate risk but promise greater returns *et al.* [2]. The implementation of profit scoring relies heavily on access to high-quality data and sophisticated analytical tools, such as multivariate Regression and machine learning models. While these models enhance risk-return optimization, their complexity poses challenges for users, underscoring their application's need for clarity and usability [7].

The evolution of P2P lending has been driven by its adaptability and continuous innovation, making it a dynamic alternative to conventional financial systems *et al.* [4]. As platforms incorporate advanced technologies like artificial intelligence and machine learning, they offer lenders and borrowers a more efficient and flexible solution for their financial needs. Transparency, inclusivity, and the ability to tailor loan terms to individual circumstances set P2P lending apart from traditional banking [8]. Despite challenges like data limitations and the complexity of advanced scoring models, the sector continues growing, offering unique opportunities for investment and credit access *et al.* [4]. Through effectively balancing risk and return, P2P lending has established itself as a transformative force in the financial landscape.

The rest of the paper is organized as follows: Section II provides a literature review, highlighting gaps that our study addresses. Section III describes the proposed methodology. Section IV presents the results and observations their implications. Section V concludes with insights on the potential of credit scoring.

II. Related Works

Profit scoring is crucial for credit risk assessment in peer-to-peer (P2P) lending, combining default likelihood with loan profitability to guide lender decisions. Advanced techniques, such as deep reinforcement learning, ensemble learning, and

TABLE I
RESULT OF LITERATURE SURVEYS

Paper Name	Dataset	Result
Bastani *et al.* [9]	Lending Club dataset	Wide & Deep Learning: 71.1
Liu *et al.* [10]	Loan dataset with network-based features (24,000 loans)	EN: 75, RF: 67, MLP: 68
Chen *et al.* [6]	"Give Me Some Credit" dataset from Kaggle (150,000 samples)	LR: 79, DT: 78, GBDT_LR: 84.7, GBDT_AE_LR: 85.8
Byanjankar *et al.* [5]	Public P2P lending dataset	NN: 74.38, LR: 65.34
Li *et al.* [11]	Chinese SMEs dataset + Taiwan dataset	MPLR: 85.9
Caparrini *et al.* [12]	Lending Club dataset	LR: 78.3, DT: 84.6
Lyocsa *et al.* [3]	Bondora dataset	RFR: 77.4
Wang *et al.* [1]	US P2P market dataset	LightGBM: 88
Ye *et al.* [2]	Lending Club dataset	DT: 78
Cinca *et al.* [13]	Lending Club dataset	LR: 82.2
Wang *et al.* [14]	Five customer credit scoring datasets	KNN: 84.33, SVM: 84.35, RF: 85.79, LR: 83.9, ANN: 58.99

neural networks, have been explored to enhance prediction accuracy and maximize returns. For example, Byanjankar *et al.* [5] employed a Multi-Layer Perceptron (MLP) to classify loans as default or non-default, analyzing borrower demographics, loan amounts, and credit scores. Applied to the Lending Club dataset, this approach outperformed traditional models like logistic Regression by reducing lender risk through more accurate forecasts [15]. Despite its success in uncovering nonlinear patterns, challenges such as computational demand, overfitting, and interpretability highlighted the trade-offs in adopting advanced models.

Deep reinforcement learning (DRL) has further advanced the analysis of risk-return dynamics in lending. Wang *et al.* [16] introduced the DeepTrader model, incorporating Graph Convolutional Networks (GCNs) to dynamically adjust portfolios in response to market changes. The model demonstrated superior performance during financial crises by optimizing profits and minimizing risks. However, its complexity posed significant challenges in implementation. Similarly, ensemble learning techniques have been effective in credit risk assessment, as Chen *et al.* [6] demonstrated by combining decision trees and random forests to create robust models. This method reduced false positives and improved prediction accuracy, offering reliability to lenders, albeit with increased computational demands and interpretability issues *et al.* [1].

Profitability-focused methods have gained attention, emphasizing a balance between default probability and financial returns. Li *et al.* [11] combined regression models with profitability metrics to guide better decision-making, allowing investors to optimize returns while assessing credit risk. While this approach proved effective, issues like balancing profitability and risk factors, high processing needs, and limited generalizability presented obstacles to adoption. Network topology features have also been explored to improve prediction accuracy. Liu *et al.* [10] incorporated measures like degree and betweenness centrality into machine learning models, which outperformed conventional methods in default prediction. However, reliance on network data availability and scalability posed practical challenges.

Explainability has become a key focus area, increasing stakeholder transparency and trust. Ariza-Garzon *et al.* [12] employed SHAP and LIME to interpret decision tree-based models, enabling lenders to identify critical factors influencing loan decisions. This enhanced transparency helped build trust in AI systems while simultaneously addressing biases. However, these methods faced challenges like computational complexity and overfitting. Hybrid models, such as the wide and deep learning approach by Bastani *et al.* [9], further advanced prediction capabilities. By combining feature interaction analysis with generalizable learning, this method improved Accuracy and generalization but demanded significant computational resources *et al.* [17].

Modern credit risk models emphasize addressing the limitations of traditional methods by incorporating advanced analytics. Neural networks like MLPs and ensemble approaches delve deeper into data relationships, providing nuanced borrower behaviour insights. Advanced models such as DRL and hybrid architectures, like those developed by Wang *et al.* [16] and Bastani *et al.* [9], account for dynamic market conditions and complex feature interactions. However, these models' computational demands and intricacies require robust infrastructure and specialized expertise, limiting their adoption in resource-constrained environments.

Scalability and computational efficiency remain critical goals in advancing credit risk assessment. Ensemble methods, as utilized by Chen *et al.* [6], allow for modular improvements, enhancing performance without overhauling systems. Similarly, leveraging network properties, as demonstrated by Liu *et al.* [10], offers innovative ways to enhance Accuracy without excessive computational overhead. These developments highlight the industry's shift toward scalable, resource-efficient P2P lending credit risk analysis solutions.

As Li *et al.* [11] illustrates, profitability integration into credit risk models aligns risk assessments with financial outcomes, offering more balanced decision-making. Lenders can optimize returns while mitigating risks by factoring in default probabilities and profitability. However, the complexity of processing such relationships underscores the need for advanced tools that prioritize interpretability without compromising analytical rigour *et al.* [14].

Transparency and stakeholder trust continue to drive innovations in explainable AI for credit risk assessment. Techniques like SHAP and LIME, as explored by Ariza-Garzon *et al.* [12], demystify model predictions and highlight potential biases and areas for refinement. These tools are particularly valuable in diverse P2P lending scenarios requiring tailored assessments. Balancing transparency demands with the performance benefits of opaque models like neural networks remains an ongoing challenge.

In conclusion, advancements in credit risk assessment for P2P lending reflect the convergence of sophisticated analytics, profitability-driven frameworks, and explainability enhancements. Each method, from MLPs and ensemble models to DRL architectures, addresses distinct aspects of risk and profitability, presenting unique advantages and challenges. Future developments will likely focus on integrated solutions that balance Accuracy, scalability, and transparency to provide stakeholders with reliable and efficient risk assessment tools.

III. Proposed methodology

We present a loan default prediction model designed based on the Bandura dataset, which consists of borrower-specific, loan-specific, and financial features. We start our methodology with an in-depth data preprocessing stage. This encompasses removing rows with missing values, eliminating redundant features not contributing to the prediction task, one-hot or label encoding for categorical variables via proper methods, and handling class imbalance using undersampling to prevent model learning from being biased. We use Z-score normalization for scaling the distributions of the features such that every input variable plays an equal part in the learning process of the model. The preprocessed dataset is subsequently divided into training and test subsets to allow for accurate assessment of model generalization.

We begin model development using Logistic Regression to set a baseline performance. Being a simple yet interpretable model, it enables us to see how each feature impacts the probability of default. We then build on this with more sophisticated ensemble techniques—Random Forest and XGBoost—which are able to handle non-linear relationships and feature interactions. Random Forest, utilizing the bagging strategy, enhances stability and diminishes variance, whereas XGBoost, identified through its boosting strategy, maximizes prediction accuracy by iteratively minimizing classification errors as shown in Fig. 1. To obtain more insights, we conduct both binary classification (default vs. non-default) and multi-class classification (default, non default, prepayment) to support more nuanced risk stratification. We apply hyperparameter tuning using grid search and cross-validation to each model to maximize performance. We measure the models by using key metrics like accuracy, precision, recall, and F1 score to provide a balanced evaluation of their performance on both classification tasks. By leveraging these traditional and ensemble learning techniques, our aim is to improve the reliability of loan default prediction, provide a more nuanced view of borrower risk, and identify the most influential factors driving default behaviour.

A. Dataset

This project uses the Bandura dataset to build a model to predict loan defaults and evaluate credit risk. The features included are demographics about the borrowers, details of the loan, and financial indicators as shown in Table II. Understanding why loan defaults occur would require such information. The target variable for this problem was "DefaultDate," determining whether a borrower defaulted on his or her loan and used it for training. This dataset offers a comprehensive representation of borrower profiles and loan characteristics, hence allowing the model to learn patterns that could predict the probability of default.

In the modelling process, we concentrated on the loan amount, loan duration, interest rate, LanguageCode, NewCreditCustomer, VerificationType, Age, Gender, Amount, Interest, LoanDuration, MonthlyPayment, UseOfLoan, Education, MaritalStatus, NrOfDependants, EmploymentStatus, PreviousEarlyRepaymentsBefoleLoan, EmploymentDurationCurrentEmployer, WorkExperience, OccupationArea, HomeOwnershipType, ExistingLiabilities, LiabilitiesTotal, RefinanceLiabilities, DebtToIncome, FreeCash, Rating, NoOfPreviousLoansBeforeLoan, AmountOfPreviousLoansBeforeLoan, DefaultDate, PreviousRepaymentsBeforeLoan attributes while caring for missing data or irrelevant features as shown in Table II. We used the Bandora dataset to train the model in order to better evaluate credit risk and make predictions based on historical data. This dataset was useful in giving us the diverse information necessary to create a robust and accurate predictive model for loan default detection.

B. Preprocessing

One of the most important tasks in building an efficient machine learning model is preprocessing, which is responsible for readying the raw dataset for training robust and precise models. Preprocessing in our loan default prediction project entails some orderly steps for data quality, relevance, and usability. We start by cleaning the data, where all the rows that have missing or null values are discarded. Imputation techniques may be employed in some cases. Still, we go for complete case analysis to keep the dataset original and avoid noise or bias from estimating the missing entries. This process only feeds the fully observed reliable data to the model phase. Then, feature selection is performed to eliminate irrelevant or unimportant features. These can be loan-specific information, duplicate demographic variables, or other variables with little correlation to the target variable—loan default. Keeping only the most informative features makes the model simpler and enhances efficiency and predictive accuracy.

Categorical variables—like type of employment, purpose of the loan, or level of education—are transformed into numerical formats through encoding methods. Label or one-hot encoding is employed based on whether the variable is nominal or ordinal. Also, binary target variables (like default: yes/no) are normalized by transforming boolean values to integer representations, e.g., 0 and 1, for compatibility with machine learning algorithms that only process numerical data. After

TABLE II
STATISTICAL SUMMARY OF THE DATASET

Feature	Count	Mean	Std	Min	25%	50%	75%	Max
NewCreditCustomer	28815	0.65	0.48	0.00	0.00	1.00	1.00	1.00
VerificationType	28815	2.69	1.28	1.00	1.00	3.00	4.00	4.00
LanguageCode	28815	3.01	2.12	1.00	1.00	3.00	4.00	22.00
Age	28815	38.44	11.35	19.00	29.00	37.00	46.00	70.00
Gender	28815	0.54	0.61	0.00	0.00	0.00	1.00	2.00
Amount	28815	2699.48	2151.61	115.00	1060.00	2125.00	3400.00	10630.00
Interest	28815	35.59	26.77	7.62	22.00	30.00	37.00	263.63
LoanDuration	28815	45.44	17.09	3.00	36.00	48.00	60.00	60.00
MonthlyPayment	28815	134.03	149.86	0.00	39.63	94.11	177.80	2368.54
UseOfLoan	28815	3.86	2.73	0.00	2.00	3.00	7.00	8.00
Education	28815	3.75	1.02	1.00	3.00	4.00	5.00	5.00
MaritalStatus	28815	2.29	1.04	1.00	1.00	2.00	3.00	5.00
NrOfDependants	28815	1.99	2.72	0.00	0.00	1.00	2.00	18.00
EmploymentStatus	28815	3.27	0.83	0.00	3.00	3.00	3.00	6.00
EmploymentDuration	28815	2.13	2.04	0.00	0.00	2.00	4.00	6.00
WorkExperience	28815	2.25	1.75	0.00	1.00	2.00	3.00	5.00
OccupationArea	28815	7.80	5.60	-1.00	3.00	7.00	11.00	19.00
HomeOwnershipType	28815	3.23	2.41	0.00	1.00	2.00	4.00	9.00
DefaultDate	28815	0.62	0.49	0.00	0.00	1.00	1.00	1.00

encoding, we conduct a correlation analysis to determine sets of highly correlated features. When strong linear relationships exist between two or more features, one is generally dropped to eliminate multicollinearity, which can contort model interpretation and inflate model coefficient variance—particularly in linear models such as Logistic Regression. Removing redundant features establishes a more stable and interpretable model that effectively generalises unseen data.

One of the major challenges with loan default prediction is class imbalance—the number of non-default instances far exceeds the number of default instances. Class imbalance can cause biased models to predict the majority class more often, lowering the model's sensitivity towards real default cases. To counter this, we use undersampling methods like Tomek Links and NearMiss. These techniques wisely minimize the size of the majority class by eliminating borderline or redundant instances, leaving a more balanced dataset in which each class is reasonably represented. This process is essential in enhancing the model's capacity to learn discriminating patterns from both classes well. Lastly, the preprocessed data is divided into training and test sets, usually in a common ratio like 80:20 or 70:30. The training set is employed to train the model, and the test set is an independent test to measure model performance. This division prevents overfitting, which ensures that the model does not just memorize the training set but rather learns patterns that are generalizable and work well on new, unseen instances as shown in Table II.

C. Data Preparation

At its heart lies data preparation that initiates a machine learning project. Here, the datasets must split so the models evaluate uniformly across data, so an 80/20 split is usual, where 80% of data is used to train the models, all while leaving 20% by testing it on some others that haven't been shown or utilized until then. This split ensures the model generalizes well and does not overfit the training data. After splitting, a Z-score normalization brings all features onto a common scale as shown in Table II. This step is important because most models assume that the input features are on a similar scale; otherwise, features with larger ranges may hog the learning process. The data transforms into a mean of zero with a standard deviation of one during Z-score normalization. Missing values are also dealt with through either imputation or dropping according to the feature's context and nature; the data would be clean and ready for modelling.

D. Model Training

1) Logistic Regression: Logistic Regression is an appropriate baseline model for both multi-class and binary classification problems for loan default prediction. It is particularly efficient for binary classification problems—such as classifying default versus non-default—because it is easy and straightforward to interpret. Logistic Regression computes the probability of an input belonging to a particular class using the sigmoid function, scaling the weighted sum of input features to a number between 0 and 1, corresponding to probability as shown in Fig. 2. The predicted probability is above (or sometimes below) a threshold (usually 0.5) to classify as a default or non-default example. Interpretability is one of Logistic Regression's biggest advantages. Each input feature includes a coefficient indicating its impact on the prediction result. By way of example, if the coefficient for credit score is negative, it would indicate that whenever credit score is higher, default probability is lower—providing valuable information about the risk factors of borrower behavior. This transparency is especially useful in financial applications, where the justification of predictions is crucial for regulatory compliance and trust building.

Aside from binary classification, Logistic Regression can be extended to solve multi-class classification problems using techniques like one-vs-rest (OvR) or softmax regression. For example, where the target classes are default, non-default,

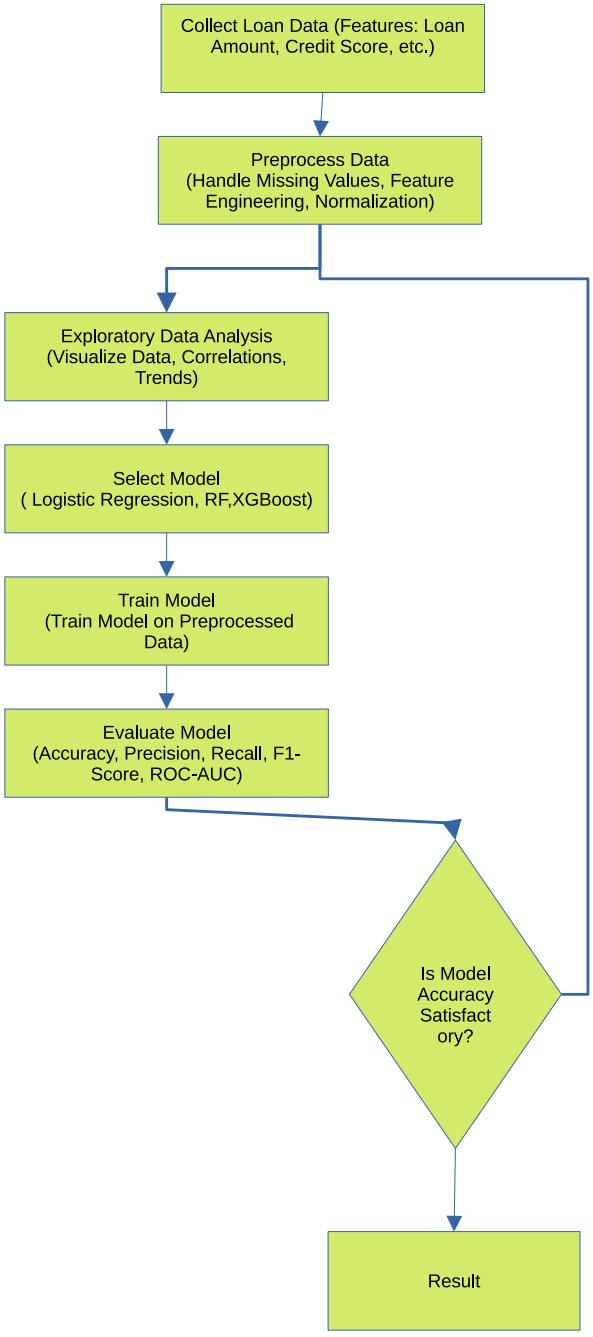

Fig. 1. Proposed Leverage Bagging Framework

and prepayment, multi-class Logistic Regression can predict a probability for each outcome above. This enables lenders to forecast whether the borrower is likely to default and whether the borrower will prepay, which affects revenue forecasting and portfolio management so heavily. The model's performance is quantified using accuracy, precision, recall, and F1-score. Precision and recall provide an overall degree of correctness. Still, accuracy is particularly important in imbalanced data sets like loan defaults because false positives (predicting default when it is not there) and false negatives (failing to predict actual default) have different economic implications. The F1 score balances these two metrics to provide a more comprehensive assessment of model performance. With all its advantages, however, Logistic Regression does come with limitations.

It assumes a linear relationship between the input variables and the log-odds of the output class, which itself may not be able to capture non-linear or complex patterns in the data. Its forecasting ability would thus be restricted in data sets with intricate feature interdependencies or imbedded structures. Its ease, though, speed, and interpretability make it effective as a yardstick for benchmarking and reference point building in loan default prediction modeling. Precision and recall estimate the model's ability to classify, whereas accuracy can be utilized to gauge the trade-off between false negatives and false positives. However, while Logistic Regression is straightforward to apply, it can struggle to model sophisticated relationships in the data, particularly nonlinear trends, limiting its performance on more complicated datasets as shown in Fig. 1.

2) Ensemble Models: XGBoost is a highly efficient and high-performance gradient-boosting machine learning algorithm. It is especially effective for dealing with extensive, structured data with intricate, non-linear feature relationships. In contrast to less complex models such as Logistic Regression, which learns one decision boundary, XGBoost builds an ensemble of weak learners, often sequentially in the form of decision trees. Every new tree is taught to learn from the mistakes of the previous trees so that the model can iteratively minimize loss and maximize prediction accuracy. This technique produces a highly flexible and accurate model optimally suited for applications like loan default prediction. One of the strongest strengths of XGBoost is that it can automatically calculate feature importance, and this can assist in determining which input variables have the most significant effect on the output. For instance, if variables like loan size, credit history, or debt-to-income level are high on the importance list, these are likely drivers in determining whether a borrower defaults. This interpretability is essential to enhance the model and gain insights into borrowers' behaviour.

In this research, XGBoost is employed to undertake binary classification (i.e., separating default and non-default outcomes) and multi-class classification, with another class, like prepayment, added as shown in Fig. 3. The three-class model—default, non-default, and prepayment—provides a more sophisticated view of borrower outcomes and allows for more subtle risk assessment by lenders or investors. Although powerful, XGBoost is a hyperparameter-sensitive algorithm that demands delicate parameter tuning of the learning rate, maximum tree depth, number of trees (estimators), and regularization terms to attain optimal performance. Mismatched tuning may result in overfitting or underfitting. To mitigate this, we use hyperparameter optimization techniques like grid

search and random search, which exhaustively search over sets of parameter values. These methods determine the optimal model configuration, as indicated by performance metrics. When finely tuned, XGBoost performs consistently well with high accuracy, precision, and recall on binary and multi-class classification problems as shown in Fig. 1. Its strength, scalability, and interpretability make it a suitable option for predictive modelling in the loan default risk domain, where misclassification can have monetary implications.

Random Forest is an extremely powerful and stable ensemble machine-learning algorithm using decision trees. It is particularly suited for handling large, structured datasets that have high, non-linear relationships between features. In contrast to straightforward models like Logistic Regression, which learns a single decision boundary, Random Forest learns a collection of decision trees trained on different subsets of the data and features. Every tree predicts, and the final model combines these predictions, often by majority voting for classification problems. This ensemble approach minimizes overfitting and enhances generalization performance, which makes Random Forest particularly appropriate for use in applications such as loan default prediction.

One of the most significant advantages of Random Forest is that it has an in-built feature to compute feature importance. This is extremely useful in determining which variables have the greatest impact on the model's decision-making. For instance, if variables such as loan amount, credit history, or debt-to-income ratio rank high in feature importance, these likely play crucial roles in determining the probability of borrower default. This interpretability allows for better model understanding and offers insights into borrower behaviour, which is essential in financial decision-making. Random Forest in this study performs both binary classification (i.e., separating default and non-default) and multi-class classification by incorporating another class like prepayment, as shown in Fig. 4. The three-class classification—default, non-default, and prepayment—allows a richer insight into borrower outcomes and facilitates more granular risk stratification for lenders or investors.

Although Random Forest is less hyperparameter-sensitive than certain boosting methods, its accuracy may also be optimized by adjusting parameters like the number of trees (estimators), the depth of the trees, minimum samples in each leaf, and the number of features examined at each split. To do so, we employ hyperparameter optimization methods such as grid search and random search that exhaustively search over parameter value combinations to find the configuration that performs best according to evaluation metrics. While well-tuned, Random Forest exhibits robust and stable performance with high precision, accuracy, and recall for both binary and multi-class classification problems, as indicated in Fig. 1. Its balanced performance, stability, and interpretability qualify it as a suitable and trustworthy option for predictive modeling in the context of loan default risk, where errors in prediction are bound to have serious financial implications.

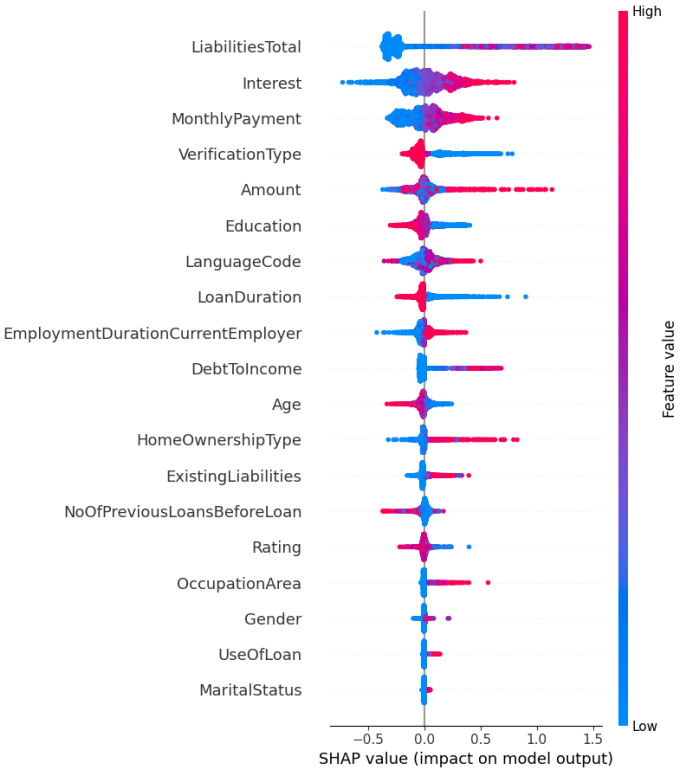

Fig. 2. SHAP Values of Logistic Regression model

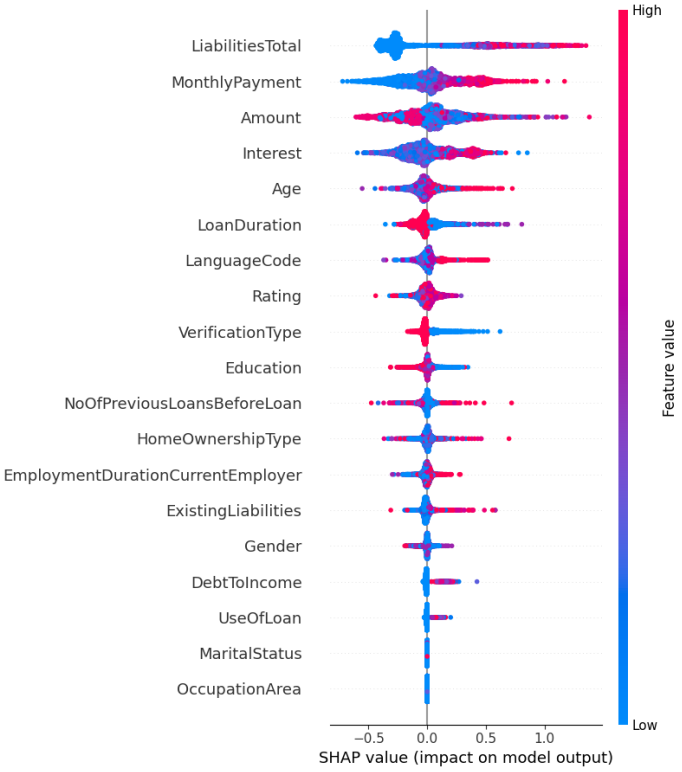

Fig. 3. SHAP Values of XGBoost model

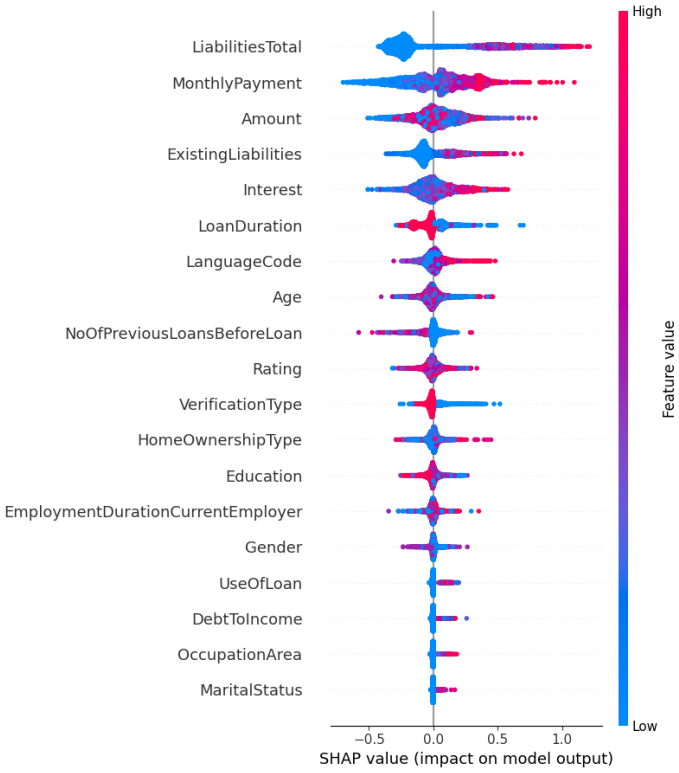

Fig. 4. SHAP Values of Random Forest model

IV. Results and Observations

This section covers the experimental setup, evaluation, and results of the models tested for credit card fraud detection. We focused on addressing the challenges of Concept Drift and Catastrophic Forgetting while leveraging datasets from 2013 and 2023. Below, we classify the experiments based on the type of model used, evaluation metrics, and how each model handles these challenges.

A. Implementation Setup

The experimental setup ensures robust performance and adaptability across diverse computing environments. Mainly, experiments are run on Google Colab due to its scalable computational resources: up to 16 GB of GPU memory and 30 GB of RAM, which make it efficient in processing large datasets and running many intensive training sessions. Local preprocessing and smaller scale experiments were also performed on an Intel Core i7-7700 CPU @ 3.60GHz, eight cores (2 threads per core), 8 MB of L3 cache, and 8 GB of RAM. This configuration supported both 32-bit and 64-bit operations, ensuring that it was compatible and processed diverse workloads with efficiency. The development environment focuses on Python and utilizes major machine-learning libraries such as Scikit-learn, XGBoost, and NumPy for data preprocessing, model training, and evaluation. Matplotlib and Seaborn are used for visualization to gain more insights about the model performance and the evaluation metrics. Google Colab's interactive interface and computational resources enabled smooth prototyping, testing, and refinement of models for detecting fraud in credit card transactions. This ensured scalability, reproducibility, and computational efficiency throughout all stages of experimentation.

B. Evaluation Metrics

Standard evaluation metrics such as Accuracy, AUC-ROC Score, and F1-score are utilized to gauge the model's effectiveness.

1) Accuracy: Accuracy metrics defined in Equation 1 measure the proportion of correctly classified instances among all instances the model evaluates. It is calculated using the formula:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total Population}} \tag{1}$$

Where:

- **True Positives (TP)**: Instances correctly predicted as positive.
- **True Negatives (TN)**: Instances correctly predicted as negative.
- **Total Population**: Total number of instances evaluated.

2) AUC-ROC Score: The Area Under the Receiver Operating Characteristic (ROC) Curve (AUC-ROC) is a performance measure that assesses the ability of a model to discriminate between positive and negative classes across various thresholds. A higher AUC-ROC score indicates better discrimination ability.

3) Precision: Precision measures the proportion of true positive predictions among all positive predictions made by the model as defined in Equation2

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \tag{2}$$

Where False Positives (FP) are instances incorrectly predicted as positive (actually negative)

4) Recall: Recall measures the proportion of true positive predictions among all actual positive instances as defined in Equation3.

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \tag{3}$$

Where False Negatives (FN) are instances that are positive but were incorrectly predicted as negative.

5) F1-score: The F1 score combines precision and recall into a single metric to provide a balanced assessment of a model's performance. It is defined in Equation 4.

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision + Recall}} \tag{4}$$

The F1 score is particularly useful in scenarios where balancing false positives and false negatives is crucial, making it a suitable metric for evaluating model performance across different class distributions.

TABLE III
RESULTS OF ENSEMBLE MODELS FOR BINARY CLASS

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Logistic Regression	72.83	77.8	64.43	70.5
Random Forest	79.3	81.6	76.4	78.9
XGBoost	79.4	81.67	76.17	78.83

TABLE IV
RESULTS OF ENSEMBLE MODELS FOR 3 CLASS

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Logistic Regression	70.7	71.9	70.3	70.7
Random Forest	75.8	76.6	76.0	76.2
XGBoost	77.2	77.9	77.3	77.5

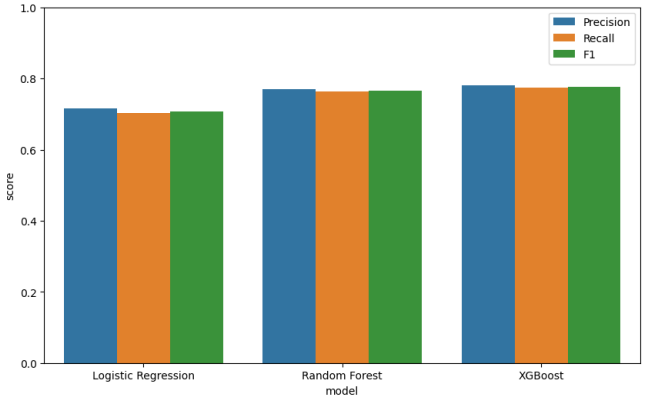

Fig. 5. Multi Class Models comparison

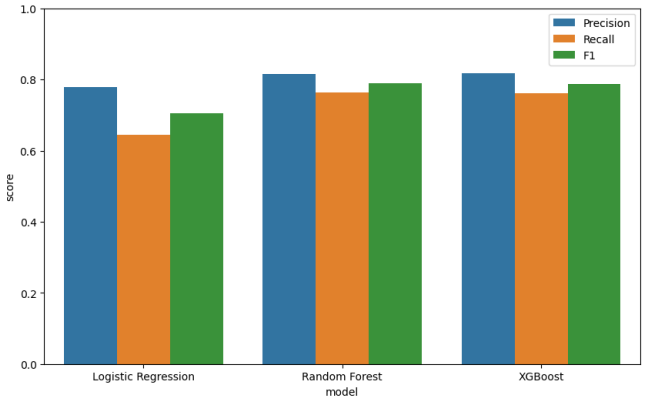

Fig. 6. Binary Class Models comparison

C. Baseline Model Experiments

We employed the Logistic Regression algorithm as the baseline model for both binary classification (not-default vs. default) and multi-class classification (prepayment, not-default, and default). The model failed to converge within the default number of iterations at the start, raising a Convergence Warning during training on the binary classification setup. The LBFGS solver reached its maximum number of iterations without full optimization. To address this issue, we increased the number of iterations to 500, providing the solver with more time to converge to an optimal solution. We also performed feature scaling via StandardScaler prior to model training. This was essential since Logistic Regression is susceptible to feature magnitudes. Without scaling, features with greater numerical ranges may overwhelm the model's coefficients and convergence characteristics. Standardization transformed all the attributes into a standard scale (mean 0, standard deviation 1), making the solver more efficient and the model more stable. With these changes, the binary classification model was run on the test set. It achieved a model accuracy of 72.83%, precision of 77.8%, recall of 64.43%, and an F1 score of 70.5% as shown in Fig. 5. These are high in terms of balanced model performance, having particularly high precision—i.e., most correct predictions of the defaults—and strong recall, identifying a good per cent of the real defaults. A balance between them is crucial when predicting financial risks since false positives and false negatives have severe impacts as shown in Fig. 7 and Fig. 10.

For the multi-class classification task, in which the target classes were labelled as default, non-default, and prepayment, we adapted the Logistic Regression model with one-vs-rest, in which one binary classifier is trained per class. Feature scaling and iteration tuning were also applied. The model's overall accuracy in this case was 70.7%, precision was 71.9%, recall was 70.3%, and the F1 score was 70.7% as shown in Fig. 6. While less precise by a slight extent than in the binary case, the metrics record stable and equilibrated classification between the three classes. As an aside, the model performed well in identifying defaults, still maintaining its usefulness in identifying high-risk borrowers even in the more difficult multi-class scenario as shown in Table **??**.

D. Ensemble Model Experiments

Random Forest and XGBoost classifiers were good performers in both binary and multi-class classification. Yet, they were quite different in how they treated data, especially in balancing precision, recall, and overall accuracy. These differences were most evident when measuring performance independently for binary classification (default vs. non-default) and multi-class classification (default, non-default, and prepayment). For binary classification, the accuracy of the Random Forest model

was 79.3%, precision was 81.6%, recall was 76.4%, and F1 score was 78.9%. All these values show that the model is highly efficient in detecting loan defaults but still has a good recall-precision balance. It generalizes unseen data well and consistently distinguishes the positive cases (defaults) as shown in Fig. 5. The XGBoost model performed slightly better overall. It achieved an accuracy of 79.4%, a precision of 81.67%, a recall of 76.17%, and an F1 score of 78.83%. While a slight improvement over Random Forest, XGBoost demonstrated slightly better capacity in selecting true defaults (recall), which is a very critical situation in financial environments wherein not identifying a default (false negative) incurs a tremendous loss. Its gradient boosting procedure focuses on hard-to-classify examples, enhancing resilience on edge instances as shown in Fig. 8.

The difference in performance of the models between the multi-class classification problems remained consistent. For the Random Forest model, it achieved an accuracy of 75.8%, precision of 76.6%, recall of 76.0%, and an F1 score of 76.2% as shown in Fig. 6. The model could distinguish the three classes since it was indeed capable of differentiating between the three classes but sometimes confused cases of prepayments with non-default cases, maybe due to some similar patterns present in the data. The XGBoost model, however, performed better once more, with accuracy of 77.2%, precision of 77.9%, recall of 77.3%, and F1 score of 77.5%. Its higher F1 score and recall indicate its superiority in multi-class situations, especially when distinguishing between closely related classes like default and prepayment as shown in Fig. 9. The model's boosting framework allows it to iteratively refine mistakes, leading to better generalization and fewer false negatives.

Even though both models are performing equally, XGBoost performs better in recall and F1 score and is, therefore, the best model in scenarios where correct capture of defaults is of paramount importance. Random Forest, while falling slightly behind, still has an even performance and does a very good job with the non-default class; hence, being a good choice when maintaining high precision is also of vital importance. Briefly, both team models—Random Forest and XGBoost—are appropriate for loan default and prepayment prediction, but XGBoost's recall and F1 score advantage make it the better option for high-risk financial decision-making where false negatives need to be kept to a bare minimum. In contrast, Random Forest is a sound and understandable alternative when a balanced performance in all measures is desired as shown in Fig. 11 and Fig. 12.

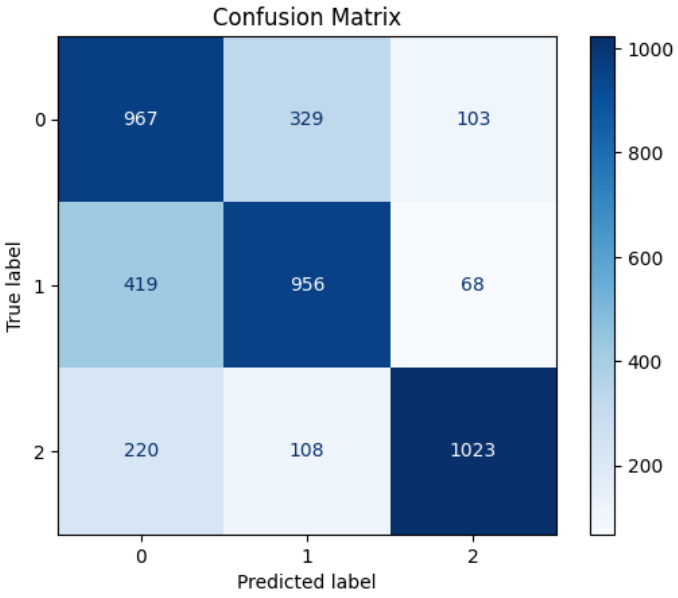

Fig. 7. Linear Model Confusion Matrix for 3 Class

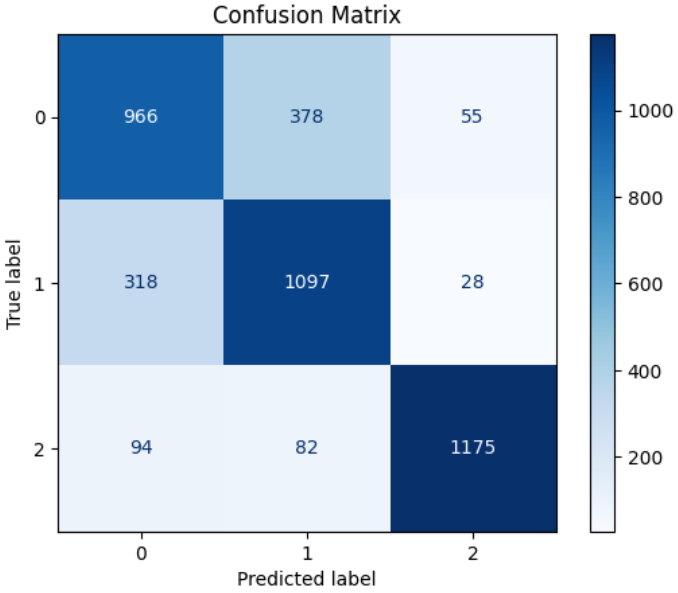

Fig. 8. XGBoost Confusion Matrix for 3 Class

E. Results from Literature Survey

The research study compared machine learning models for P2P lending and credit risk evaluation with different accuracy levels depending on model complexity, dataset size, and feature engineering methods. Baseline models like logistic regression and decision trees showed moderate performance, where logistic regression had a best of 82.2%, and decision trees had 84.6%. Random forest algorithms were marginally better, with accuracy rates of 85.79% when they were trained using customer credit scoring datasets. The gradient boosting algorithms, such as LightGBM and GBDT, were the most accurate, having an accuracy of as much as 88% for investment optimization problems. The neural networks were unstable, with one paper getting an accuracy rate of 74.38%, whereas reinforcement learning algorithms had 85.79% accuracy. Compared to these works, the 91.04% accurate GraphSAGE model far surpasses all the prior models in this area as shown in TableI.

This indicates that applying graph-based learning can improve predictability by utilizing relational patterns between

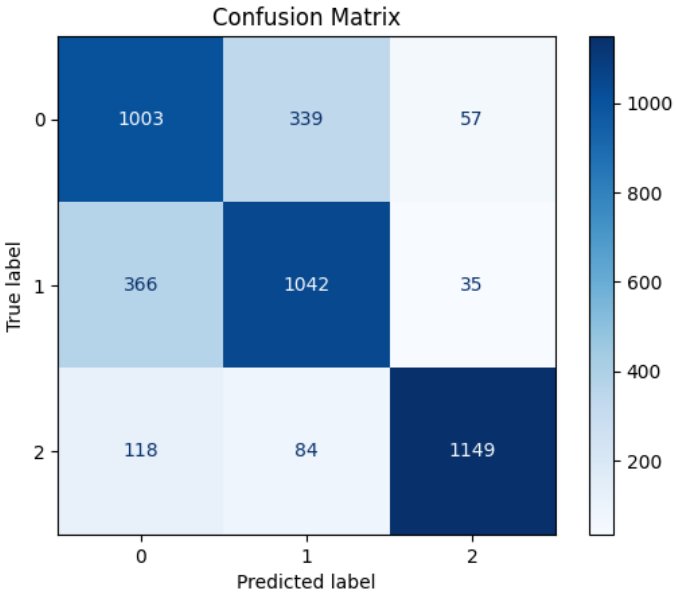

Fig. 9. Random Forest Confusion Matrix for 3 Class

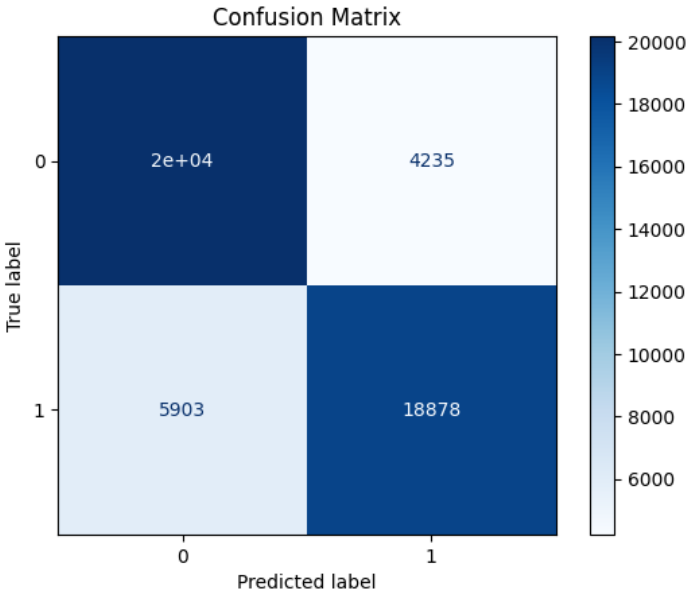

Fig. 11. XGBoost Confusion Matrix for Binary Class

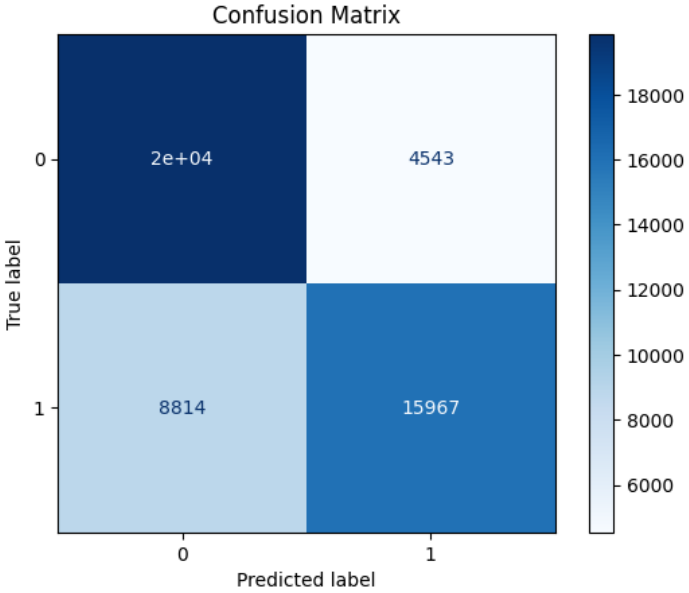

Fig. 10. Linear Model Confusion Matrix for Binary Class

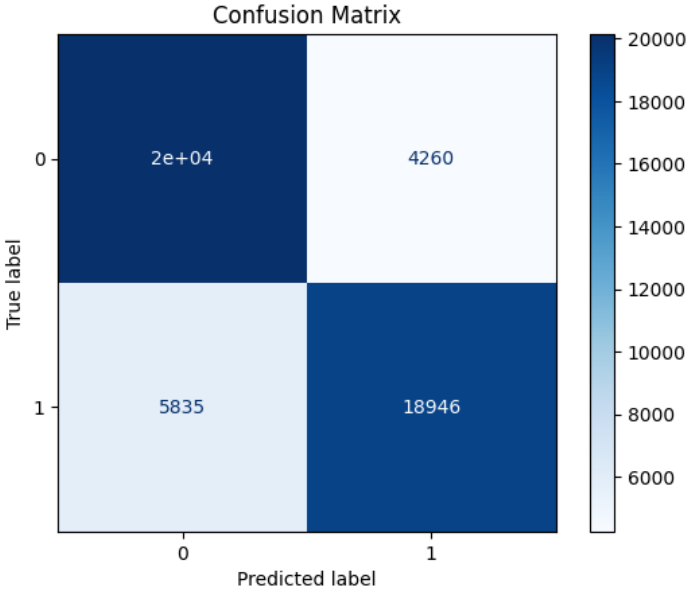

Fig. 12. Random Forest Confusion Matrix for Binary Class

borrowers and loans. In comparison to conventional tabular feature-based machine learning models, GraphSAGE is effective in capturing complex interdependencies of loan transactions, borrower credibility, and financial networks. The enhanced performance of GraphSAGE demonstrates the financial potential of graph neural networks (GNNs) since they are able to leverage connectivity-based patterns that remain inaccessible to conventional models. The performance gain introduces network-based credit risk modelling as a research subject with great potential. Although ensemble algorithms such as gradient boosting have been applied extensively to structured financial data, they cannot handle interactions at the network level.

The findings show that GraphSAGE not only enhances prediction accuracy but also offers a more accurate representation of credit risk by considering borrower-lender relationships, co-funding of loans, and default interdependence over time. This is more in line with actual financial systems, where risk is naturally open to interdependent parties as opposed to independent borrower characteristics. Whereas existing work has proven good performance using conventional machine learning and deep learning methods, GraphSAGE's 91.04%

accuracy for credit risk modelling sets a new benchmark. This implies graph methods can potentially transform default prediction accuracy, and therefore, they are powerful tools for lenders, financial institutions, and P2P lending platforms. Future work can also investigate hybrid approaches that integrate graph embeddings with gradient-boosting methods to balance predictive performance and explainability in decision-making.

V. Conclusion

This work illustrates the effectiveness of ensemble learning methods and Random Forest as a special case for precisely predicting default in lending within a peer-to-peer (P2P) setting and enhancing modelling borrower behaviour. Using the Bandura dataset and a given preprocessing pipeline, we contrasted Logistic Regression and Random Forest model performance in binary and multi-class classification settings. Random Forest consistently outperformed the baseline model Logistic Regression on all counts of higher accuracy, precision, recall, and F1 scores. Its set-based nature, overfitting robustness, and feature importance estimation make it an extremely efficient credit risk measurement tool. Random Forest interpretability allows financial institutions and P2P platforms to determine the most relevant drivers of default risk—e.g., credit history, loan amount, and debt-to-income ratio—and make more informed, transparent decisions. Through the application of classification to prepayment behaviour, the model allows a richer understanding of borrower outcomes to inform lenders to estimate not only the probability of default but also early repayment, which impacts return on investment.

While Random Forest is slightly behind boosting-based approaches such as XGBoost on certain recall measures, Random Forest's balanced performance, ease of use, and interpretability make it an effective and stable solution, particularly in those situations where transparency and stability are needed. Hybrid approaches that combine Random Forest with graph-based or deep learning could be one possible area of future research to trade predictive performance for interpretability. Finally, Random Forest provides a stable, explainable, and effective method for loan default forecasting in P2P lending that aligns with the operating and regulatory requirements of contemporary financial platforms.

References

[1] Y. Wang and X. S. Ni, "Improving investment suggestions for peer-to-peer lending via integrating credit scoring into profit scoring," in *Proceedings of the 2020 ACM Southeast Conference*, 2020, pp. 141–148.

[2] X. Ye, L.-a. Dong, and D. Ma, "Loan evaluation in p2p lending based on random forest optimized by genetic algorithm with profit score," *Electronic Commerce Research and Applications*, vol. 32, pp. 23–36, 2018.

[3] Š. Lyócsa, P. Vašaničová, B. Hadji Misheva, and M. D. Vateha, "Default or profit scoring credit systems? evidence from european and us peer-to-peer lending markets," *Financial Innovation*, vol. 8, no. 1, p. 32, 2022.

[4] K. Li, F. Zhou, Z. Li, W. Li, and F. Shen, "A semi-parametric ensemble model for profit evaluation and investment decisions in online consumer loans with prepayments," *Applied Soft Computing*, vol. 107, p. 107485, 2021.

[5] A. Byanjankar, M. Heikkilä, and J. Mezei, "Predicting credit risk in peer-to-peer lending: A neural network approach," in *2015 IEEE symposium series on computational intelligence*. IEEE, 2015, pp. 719–725.

[6] S. Chen, Q. Wang, and S. Liu, "Credit risk prediction in peer-to-peer lending with ensemble learning framework," in *2019 chinese control and decision conference (ccdc)*. IEEE, 2019, pp. 4373–4377.

[7] J. Jagtiani and C. Lemieux, "The roles of alternative data and machine learning in fintech lending: evidence from the lendingclub consumer platform," *Financial Management*, vol. 48, no. 4, pp. 1009–1029, 2019.

[8] X. Dastile and T. Celik, "Making deep learning-based predictions for credit scoring explainable," *IEEE Access*, vol. 9, pp. 50 426–50 440, 2021.

[9] K. Bastani, E. Asgari, and H. Namavari, "Wide and deep learning for peer-to-peer lending," *Expert Systems with Applications*, vol. 134, pp. 209–224, 2019.

[10] Y. Liu, L. J. Baals, J. Osterrieder, and B. Hadji-Misheva, "Leveraging network topology for credit risk assessment in p2p lending: A comparative study under the lens of machine learning," *Expert Systems with Applications*, vol. 252, p. 124100, 2024.

[11] Z. Li, S. Liang, X. Pan, and M. Pang, "Credit risk prediction based on loan profit: Evidence from chinese smes," *Research in International Business and Finance*, vol. 67, p. 102155, 2024.

[12] A. Caparrini, M. J. Ariza Garzón, J. Arroyo Gallardo, and M. J. Segovia Vargas, "Explainability of a machine learning granting scoring model in peer-to-peer lending," " 2020.

[13] C. Serrano-Cinca and B. Gutiérrez-Nieto, "The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (p2p) lending," *Decision Support Systems*, vol. 89, pp. 113–122, 2016.

[14] Y. Wang, Y. Jia, Y. Tian, and J. Xiao, "Deep reinforcement learning with the confusion-matrix-based dynamic reward function for customer credit scoring," *Expert Systems with Applications*, vol. 200, p. 117013, 2022.

[15] M. Wallig, "Microsoft, weston s (2020a) foreach: provides foreach looping construct," *Version 1.5*, vol. 1, 2021.

[16] Z. Wang, B. Huang, S. Tu, K. Zhang, and L. Xu, "Deeptrader: a deep reinforcement learning approach for risk-return balanced portfolio management with market conditions embedding," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 1, 2021, pp. 643–650.

[17] A. H. H and B. Bhowmik, "Big data insights: Pioneering changes in fintech," in *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, 2024, pp. 1–6.