PEER-TO-PEER LOAN DEFAULT PROPHECY IN FINTECH: A COMPARATIVE ANALYSIS OF THE PREDICTIVE PERFORMANCE OF MACHINE LEARNING MODELS

Amandeep Singh[1]

Abstract

This study investigates the predictive accuracy of four machine learning methods in forecasting loan defaults within the peer-to-peer (P2P) lending sector of FinTech, for deeper penetration in the digital economy, primarily due to technological and business advancements. Unlike previous studies, which focuses only on removing data imbalances through only oversampling technique, our research incorporates both oversampled and under-sampled balanced dataset to show the effect on performance metrics. The study utilizes logistic regression, random forest, decision tree, and XGBOOST, using data from 8, 41,476 observations from Lending Club (LC) across fifteen variables. To overcome the imbalances in the dataset, the random oversampling example technique (ROSE) is applied. The analysis exhibits that the XGBOOST model outperforms logistic regression, random forest, and decision tree models on the original and balanced dataset (under sampled dataset) in terms of accuracy, specificity, and sensitivity, but its performance remains suboptimal. However, on balanced datasets (oversampled), the random forest model surpasses XGBOOST, decision tree, and logistic regression in accuracy, sensitivity, and area under the receiver operating characteristic curve. Logistic regression and decision tree results are suboptimal. Therefore, banks and FinTech's employing the random forest model on a balanced (oversampled) dataset can lead to improved prediction results, minimizing default risks for P2P platforms. The study reveals that the predictive ability of models can help FinTech firms where borrowers are predominantly from the low-income population bracket. Even a slight improvement in predictive performance can save FinTech from a million-dollar loss, which will ensure financial stability in the P2P lending sector.

Keywords: FinTech P2P lending, Machine Learning, Random Forest, Logistic Regression, XGBOOST, Default- prediction

JEL Classification: O32, G23, C45

[1] Assistant Professor, Commerce , Sri Guru Tegh Bahadur Khalsa College, University of Delhi

1. Introduction

Peer-to-Peer (P2P) lending, a form of FinTech, commenced with the launch of Zopa in 2005 in the U.K. After this, Prosper and Lending Club started their P2P lending services in the USA. While, the concept of P2P lending is not novel historically, individuals used to borrow and lend money to one another these modern P2P platforms have formalized the market. Going back to 2008, a crucial year in the first decade of the 21st century, manifested by the global financial crisis that trembled the world, due to which banks have witnessed a high degree of regulatory compliance in the past decade, which has created a competitive disadvantage(Amstad, 2019). Increasing regulatory burdens and high capital requirements have enabled regulatory arbitrage opportunity for FinTech. Furthermore, AI-driven technology has disrupted markets, providing better service experience, convenience, and customer outreach for financial services (Buchak et al., 2018; Ryu & Chang, 2018). The major reasons behind AI growth are ease of loan applications, the quick turnaround time for loan disbursal, efficient sales management, and better experience for customers, no lock-in period, and prepayment charges (Anil & Misra, 2022). FinTech's emergence reduces incumbent banks' profits, prompting risky behavior. Regulators respond with stringent rules, spurring banks to explore shadow banking, evading regulations. Post-subprime crisis, U.S. regulations tightened, fueling shadow banking growth. Shadow banks' mortgage market share in the U.S. nearly tripled from 2007 to 2015 (Vives, 2020).

P2P market experienced growth but also face mixed opinions, hailed as innovation yet criticized as Ponzi schemes (J. J. Xu, 2016). Lack of supervision resulted in increased credit risk, with over 800 platforms collapsing due to fraud, therefore, credit risk assessment crucial for to maintain financial stability in order (Guo, 2020). The factors posing potential financial risks for P2P Lending platforms are unreal investments, platform revocation, and borrowers' inability to repay loans (J. Xu et al., 2021). Machine Learning (ML) algorithms identify low-risk borrowers and price credit more accurately, which is a major divergence from traditional banking (Vallee & Zeng, 2019). Balancing the benefits and challenges of AI it is critical for effective integration and governance in diverse business environments. With the rise of big data and Internet of Things, machine learning has been widely used in the area of computational finance such as credit scoring and algorithm trading(Appio et al., n.d.; Johnson et al., 2022). Hence, this study compares logistic regression, random forest, decision tree, and XGBOOST models for loan default prediction, concentrating on accuracy, specificity, sensitivity, kappa, balanced accuracy, and AUC. It

accentuates recognizing key aspects distinguishing good and bad borrowers. Unlike previous studies, our research focuses on a comprehensive comparison of these four ML techniques on both imbalanced and balanced datasets (oversampled and under sampled). The present study aims to provide an efficient and reliable loan default prediction method to mitigate default risk and ensure the financial stability of P2P platforms.

2. Literature Review

Recent studies reveal the rise in the utilization of machine learning usage for loan default prediction, replacing traditional statistical models. With the large volumes of data and complexity in feature dimensions, together with ongoing advancements in data-driven methodologies, have facilitated the adoption of more sophisticated predictive models. The advanced techniques such as Support Vector Machines (SVM), Random Forest (RF), Artificial Neural Networks (ANN), and Gradient Boosting Method show potential for enhancing the accuracy of loan default prediction (Moula et al., 2017).

For instance, Emekter et al. (Emekter et al., 2015) applied a logistic regression technique to predict the probability of borrowers' default and established that the revolving line utilization, debt-to-income ratio, FICO score, and credit grade are vital factors. Random Forest and logistic regression were used by Ma et al. and Coser et al. (Coşer et al., 2019; Ma et al., 2018) to find a series of prediction models for evaluating the probability of a customer's loan default. The problems associated with logistic regression due to high-dimensionality were identified when the independent variables are too many it blurs the result (Vallee & Zeng, 2019). A study conducted employed the oversampling technique SMOTE but overlooked the complementary under sampling approach for achieving data balance (J. Xu et al., 2021). Additionally, research utilized ML techniques, including gradient boosting, light GBM, and XGBoost, on an imbalanced dataset, with an over-emphasis on reporting accuracy, which highlights that other performance measures are given less importance (Zhou et al., 2019) . The study applied wrapper feature selection method to identify the optimal features affecting the defaults in P2P lending (Sam'an et al., 2024).

The empirical study employed logistic regression and Neural Networks (NN) to forecast firm's loan defaults, finding that NN outperformed traditional logistic regression based on financial ratios, tax arrears, annual reports, and submission delays (Kohv & Lukason, 2021). Correspondingly, it was found the AdaBoost model outperformed Random Forest, XGBOOST, K Nearest Neighbor, and Multilayer Perceptron in predictive accuracy using real datasets (Lai, 2020).

Adding further, applied adaptive synthetic sampling to deal with data imbalance, establishing that their fusion model outperformed logistic regression, random forest, and cat boost (Li et al., 2021). XGBOOST model demonstrated superior predictive performance when compared to logistic regression (Lubis et al., 2024).

It has been established that the random forest (RF) model, with 80% accuracy, outperformed the decision tree (DT) model, which attained 73% accuracy (Madaan et al., 2021). However, regardless of its higher accuracy, RF misclassified some non-defaulters as defaulters. Existing literature proposed that the decision tree model based on short-term credit assessment integrated SMOTE to improve the performance of DT model on imbalanced dataset (Chang et al., 2016). The precision and recall rate are higher than that of logistic regression model. Authors utilized DT model on imbalanced and balanced under sampled dataset and concluded that predictive performance of a DT model based on a balanced data set is more reasonable compared than that of imbalanced data set (Syed Nor et al., 2019). Adding further, study explained that identification of necessary features affecting the default may mitigate the risk involved in lending and utilized Boruta and mRMR to identify the important features (Hegde et al., 2023). Prior research presented a novel LGB-XGBOOST stacking model, outstripping individual models such as XGBOOST, Light GBM, and CatBoost, with a 24% higher recall rate and 6.71% higher AUC (Liu et al., 2022). Furthermore, random forest was employed to analyze how loan features like gender, education, employment type, business type, loan term, and marital status impacts bank loan application decisions (Dansana et al., 2024).

Table 1: Research gaps in existing literature

Author	Methodology	Findings	Research Gaps
Emekter et al. (2015)	Logistic Regression	credit grade, debt-to-income ratio, FICO score and revolving line are important in determining loan default	Need to address the issues related to accuracy, sensitivity and specificity of the model.
Ma et al. (2018)	LightGBM & XGBOOST	The predictive performance of LightGBM is better than XGBOOST.	Data imbalance issue was not addressed which leads to error rate of 19.90%.
Coser et al. (2019)	LightGBM, XGBOOST, Logistic	The AUC of random forest model was 0.89, showing best result	Utilizing SMOTE or ROSE technique could have produced far better results.

	regression and Random Forest (RF)		
Xu et al. (2021)	Random Forest (RF), XGBOOST, GBM, Neural Network (NN)	The mobile phone, video, education, job and income verification plays significant role in determining the loan defaults. The predictive performance of RF is superior compared other models.	Utilized SMOTE to generate oversamples. While the SMOTE is capable of performing both under sampling and over sampling.
Dansana et al. (2023)	Random Forest (RF)	Analyze the loan approval on different aspects such as gender, educational qualifications, type of employment, type of business, loan term, and marital status.	The study didn't discuss about the predictive performance of RF based on performance metrics.
Zhu et al. (2023)	Logistic Regression, Decision Tree, XGBOOST and Light GBM.	The XGBOOST and LightGBM outperform LR and DT in terms of predictive ability	Imbalances present in the large dataset were not addressed using ROSE or SMOTE method.

Source: Author's compilation

From the snippets placed in table 1, it is apparent that existing literature primarily compares imbalanced datasets with oversampled ones, with a very few studies examining under sampled balanced datasets. Further, SMOTE technique is commonly used to tackle data imbalances, while some studies proposes stacking models to enhance predictive performance.

3. Research Objectives

The in-depth review of existing studies exhibits a literature gap in applying SMOTE or ROSE to address dataset imbalances. While studies did not utilize SMOTE or ROSE (Coşer et al., 2019; Ma et al., 2018; X. Zhu et al., 2023), one study (J. Xu et al., 2021) only applied oversampling techniques, neglecting under sampling methods to tackle dataset imbalances. The present study aims to address this gap by evaluating the effectiveness of SMOTE, ROSE, and combination of oversampling and under sampling techniques in handling imbalances.

The Research objectives are:

- To analyze the performance of machine learning models namely logistic regression, random forest, decision tree and XGBOOST model for loan default prediction of P2P lending platforms for imbalanced dataset and balanced dataset.
- To compare the predictive efficacy of all four models on imbalanced and balanced dataset.

The primary goal of this paper is to present an effective loan default prediction technique to mitigate default risk and ensure stability in P2P lending platforms. Unati et,al (Uddin et al., 2023) emphasized that identifying suitable borrowers is challenging for banks, as loan acceptance affects the revenue. Improvement in loan default prediction accuracy directly impacts platform profitability (Coşer et al., 2019). By providing such a methodology, this study seeks to enhance the reliability and efficiency of prediction systems, essential for maintaining stakeholder's trust and ensuring the sustainable growth of P2P lending platforms.

4. Research Methodology

Machine learning is a branch of computer science that focuses on giving computers the capability to learn. The objective is to create an algorithm that can learn from the data and make predictions about the data. With the surge in loan applications as a large portion of the population applies for loans from banks and non-banking financial institutions accurate assessment of loan application becomes more challenging due to rising default rates. The fundamental question that needs to be answered is how much risk is associated with each borrower. The attributes of the borrowers are very crucial for predicting the default risk; however, the selection of the model should be such that can predict the borrowers' default with a high level of accuracy. Jordan and Mitchell (Jordan & Mitchell, 2015) explained that decreased cost of computation and online availability of data are the major drivers of growth in machine learning.

1. Logistic Regression: Logistic Regression is widely regarded as the preeminent supervised machine learning algorithm. It is utilized to forecast categorical dependent variables based on a specific collection of independent factors. The dependent variable should possess categorical or discrete values. The value can be either non-default or default, represented by 0 or 1 respectively. The output will not be a precise 0 or 1 value; instead, it will assign a probabilistic value ranging between 0 and 1. We define a random variable Y with binomial distribution, i.e.,$Y: \Omega \rightarrow 0,1$ on probability space Ω, F, P Logistic regression is a method in which the Y response variable has a binomial distribution. The logistic regression model predicts the likelihood of the dependent variable Y based on the observed input variables X (Rymarczyk et al., 2019). It is employed to resolve categorization difficulties. The loan default probability (success) is calculated by $PY = 1|X$, while the chance of no default (defeat) is $PY = 0|X$. A logistic regression model computes the likelihood of belonging to one of the two groups in the dataset:

$$PX,a=11+e-a.x$$

Whereas $(0|X, \alpha) = 1 - (1|X, \alpha)$. The odds ratio is represented by P1-Pwhere $P = probabilityofdefaultand\ 1 - Pisprobabilityofnodefault$. In 2018, (Nalić & Švraka, 2018) explained that simple to understand, sturdy performance, and easy implementation are the primary advantages and reasons for its widespread application.

2. Random Forest: Random Forest is a supervised machine learning algorithm constructed from a decision tree algorithm, commonly used for regression and classification problems. It creates multiple decision trees during training phase and predicts outcomes based on the mode of the classes or the average prediction of the individual trees (L. Zhu et al., 2019). A random forest is a collection of independent decision trees. It is preferred for several reasons: it handles large datasets efficiently and surpasses other algorithms in accuracy, manages errors in imbalanced datasets effectively, and maintains high precision even with significant missing data. Random Forest is useful for evaluating loan applicants' creditworthiness, predicting loan default probabilities, and forecasting customer preferences based on historical behavior (L. Zhu et al., 2019).

3. Decision Tree: A decision tree is a type of supervised machine learning algorithm that may be used for both classification and regression tasks. It is capable of predicting a model and drawing conclusions based on a given set of observations. The decision tree is a classification tree if the goal variable is 1 or 0, and a regression tree if it is continuous or real. To homogenize data, a decision tree algorithm binary separates feature space into subsets. The name decision trees come from their tree-like implementation (Granström & Abrahamsson, 2019). Graphs are utilized for probability analysis. In a tree design, core nodes show borrower qualities, branches show test results, and leaf nodes show categories (X. Zhu et al., 2023).

4. XGBOOST: XGBOOST, which stands for Extreme Gradient Boosting, is a powerful and widely used machine learning algorithm known for its efficiency and effectiveness in various types of data science competitions and real-world applications. It was developed by Tianqi Chen. XGBOOST is an improvement over the decision tree and gradient boosting algorithm. Through a large number of iterations, each iteration produces a weak classifier, and each weak classifier is trained on the bias of the result of the previous classifier. XGBOOST is an ensemble learning algorithm based on gradient boosting with decision trees as base learners. XGBOOST incorporates regularization techniques to control the complexity of the individual trees and prevent overfitting. XGBOOST is designed for efficient parallel and distributed computing, making it scalable to large datasets.

$$\hat{y}i=\emptyset xi=K=1Kfk(xi)$$

Where is fx the regression tree, K is the number of regression trees, fk(xi) x is the score of the i-th, observation given by the k-th tree. XGBoost is an efficient and widely used machine learning algorithm (Li et al., 2021). They applied the XGBOOST technique for P2P loan default prediction and it performed well in predicting the outcome.

4.1 Data Source

The Lending club dataset retrieved from the Kaggle website covers the loan disbursement from 2012 to 2015, with loan terms of either 36 months or 60 months (Ferozi, 2018). Therefore, loans disbursed in 2015 with a 36-month term would mature in 2018, while those with a 60-month term would mature in 2020. The post loan performance tenure is essential to get the model trained on actual outcome. The choice of the dataset period is determined by the presence of high-quality enhanced data and the requirement for the conclusions to be comparable. Authors (Madaan et al., 2021), analyzed Lending Club data from 2007 to 2015, adding context to our dataset selection, particularly for tracking actual loan consequences post-term. Moreover, Muslim et al. (Muslim et al., 2023) also obtained a Kaggle Lending Club dataset spanning 2007-2015 with a sample size of 5600. Our study dataset comprises of 8,41,476 observations and 30 variables. Out of these characteristics, specific ones such as borrower id, issue date, and end date were deemed irrelevant, while several variables have been determined to be superfluous in model inclusion. Thirteen variables were found to be redundant for the study, only sixteen predictor variables and one dependent variable were considered for the study. The dependent variable is the loan condition, which can be classified as either default (1) or no default (0), contingent upon whether the borrower fails to repay the loan or successfully repays it, respectively.

4.2. Data imbalances

Out of the 841,476 observations, approximately 92% of the observations are classified as excellent loans, while the other 8% are classified as poor loans. With loan default observations accounting for only 8% of the dataset, there is a significant issue of data imbalance. To address this, the Synthetic Minority Oversampling Technique (SMOTE) can be applied, the imbalance negatively affects the predictive accuracy of the model (Wang, 2022). The oversampling technique entails generating additional instances of the minority class by randomly duplicating the original minority samples. This process serves to augment the number of minority samples and achieve a more balanced distribution among the various classes. While SMOTE is a widely used method for

addressing class imbalances, it is not without its constraints(Blagus & Lusa, 2013; Jiang et al., 2022). SMOTE faces significant challenges, including issues with overfitting, interference from noise, and limitations in selecting appropriate neighbors. To mitigate the issue of overfitting, one might employ feature engineering techniques such as lowering dimensionality and emphasizing crucial information, increasing the size of the training dataset, implementing early stopping, utilizing ensemble approaches.

4.3. Software Selection

We have chosen the R programming package for data analysis due to its unique benefits. The software allows for seamless incorporation of fundamental packages like 'Caret', 'Random forest', and 'Boruta', which offer powerful tools for data preparation and machine learning. In addition, R is a widely utilized statistical software that provides specialized packages for advanced data analysis and visualization.

4.4. Data Description

The wrapper method for feature selection was employed to distinguish important and unimportant variables. Out of 16 predictor variables, the wrapper method identified 15 as important and only one as unimportant. The unimportant variable related to the application type category indicates whether the loan was applied for by an individual or a joint borrower. The *Boruta* package in R was subsequently applied to the dataset. Due to resource constraints with a big dataset, a subset with n=50,000 and a maximum of runs equals to 30 was created.

Table 2: Description of Predictor Variables Considered for the Study

Sno.	Variable	Description
1.	**Loan amount**	Amount of loan applied by the borrower
2.	**Employment length**	The length of employment is in years. 0 means less than 1 year of service and 10 means 10 and more years of employment. So the value is between 0 and 10.
3.	**Home ownership**	Home ownership status is information provided by the borrower during registration or obtained from the credit report. We define three binary variables that identify whether a borrower has a mortgage, is a rent, or owns their home/other situation. 1 means rent, 2 means own and 3 means a mortgage.
4.	**Income category**	If the annual income is less than $100000 then the income category is low, if the annual income is between $100000 and $200000 then the income category is medium, and if the annual income is above $200000 then the income category is high. The self-reported annual income in US dollars provided by the borrower during registration.

5.	**Term category**	The loan term is either 36 months or 60 months. 36 months loan's term category is 1- and 60-months loan's term category is 2
6.	**DTI**	A ratio calculated using the borrower's total monthly debt payments on total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-report monthly income.
7.	**Annual Income**	Annual income reported by the borrowers
8.	**Purpose Cat**	It shows the purpose of a loan
9.	**Interest payment Category**	Interest payment categories are low, medium, high
10.	**Interest rate**	It shows the rate of interest charged from the borrower
11.	**Grade Category**	Grade category A represents a higher likelihood of repayment and F represents a higher likelihood of non-payment
12.	**Total Payments**	It shows the total amount paid by the borrower
13.	**Total recovery of Principal**	It shows the total recovery of the principal amount
14.	**Recoveries made**	Recoveries represent the funds that the lender has successfully collected from borrowers after they have defaulted on their loans
15.	**Installment**	Installments paid

Source: Authors' compilation

5. Data Analysis

5.1 Descriptive Statistics

The descriptive statistics placed in table 3a infers an average annual income of the borrower was $75,320, the median income represent that 50% of the borrowers earn less than equal to $65,000 with a standard deviation of $64769. Employment length averages 6.10 years, with a median of 6.05 years and a standard deviation of 3.50 years. Interest rates range from 5.32% to 28.99%, with an average of 13.30% and a median of 12.99%, displaying significant variation with standard deviation of interest rates is 4.40%. The average debt-to-income ratio is 18.4%, suggesting that this percentage of monthly gross income goes to debt payments. Ideally, this ratio falls between 36% and 43%, which the lending platform considers, with only 74 borrowers above 43%.

Table 3a: Descriptive statistics of Interval variable

Description	Minimum	Maximum	Mean	Median	Standard Deviation
No. of Observations = 841476					

Annual Income ($)	0	9500000	75320.16	65000	64769.1
Employment length (in Years)	0.5	10	6.10	6.05	3.50
Interest rate (%)	5.32	28.99	13.30	12.99	4.40
Debt to Income ratio (%)	0	9999	18.40	17.89	17.54

Source: Authors' compilations

The home ownership status of borrowers is evident from table 3b, it is interesting to note that a majority (over 50%) have a mortgage, approximately 10% own their homes outright, and around 40% are renters, hence, use of leverage is prevalent.

Table 3b: Descriptive statistics of categorical variable

Description	Proportion
Loan condition	
No Default	92%
Default	8%
Home Ownership Status	
Mortgage	50.25%
Renter	39.77%
Home owners	9.98%
Other	0.02%
Loan Term	
36 Months	69.77%
60 Months	30.23%
Income Category	
Low-Income Cat	82.05%
Middle-Income Cat	16.04%
High-Income Cat	1.91%
Loan Grades	
A	16.34%
B	28.66%
C	28.06%
D	15.80%
E	7.97%
F	3.17%
Interest payment Type	
High	48%
Low	52%
Loan Purpose	
Debt consolidation	59.71%
Credit Card	23.74%
House Improvement	5.76%
Various Purposes	10.79%

Source**:** Authors' compilations

The categories ANY, NONE, and OTHER make up insignificant proportions and can be ignored in the analysis. Most borrowers, about 69.8%, opt for a 36-month loan term variable. According to the data, a significant portion of borrowers on the Peer-to-Peer Lending platform have a loan term of 60 months, with 30.2% falling into this category. A small percentage of borrowers, less than 2%, belong to the high-income category, while 16% belong to the medium-income category.

The majority, accounting for 82% of borrowers, fall into the low-income category. When it comes to loan grades, the breakdown is as follows: 16.3% fall under grade A, 28.6% under grade B, 28% under grade C, 15.8% under grade D, 8% under grade E, and 3.17% under grade F. Regarding interest payment categories, 48% of the loans have high-interest payments, while 52% have low-interest payments. Many borrowers have a primary goal of debt consolidation, with credit card usage being a close second. Around 6% of borrowers choose home improvement loans, while 10% opt for loans for various purposes such as education, vacations, weddings, renewable energy, car purchases, and more. Overall the data reflects the scope of increase in short as well as long term credit. In such scenario, asserting the servicing of loan on time is the most important objective of the FinTech's, which is gauge able by these machine learning models. This study deciphers its accuracy, reliability and sensitivity as well.

5.2 Data Preprocessing

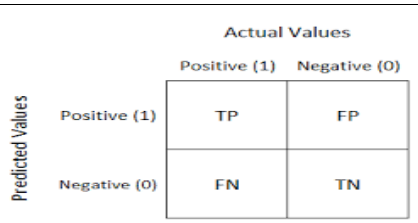
Subsequent to descriptive analysis, to preprocess a dataset to make it suitable for modeling, data processing techniques involve the substitution of missing values, conversion of variable formats, removal of redundant variables, standardization of data, selection of the most pertinent variables for modeling, and division of the dataset. The dataset is flawless, with no missing data, and consists of clean secondary data. String variables have been discovered and converted into numeric values. In addition, specific categorical variables have been converted into numerical representations, such as the values 1 or 2. To enhance the algorithm's effectiveness during the training phase, numerical variables have been standardized.

During the process of data pre-processing, it is crucial to divide the data into two distinct groups: one group is used for training the model, while the other is used for testing. The two sets, commonly referred to as the training and test data, comprise 80% and 20% of the data, respectively. Although there is no explicit guideline in the literature, it is generally advised to have a training set that is bigger than the testing set in order to achieve optimal model training and evaluation.

5.3. Evaluation Indicators

To assess the success of the machine learning model, many performance measures can be used. The performance metrics assist in determining how effectively a model works on a given dataset. The following indicators were used to evaluate the performance of the logistic regression, random forest, decision tree and XGBOOST model:

Table 4: Description of Performance Metrics

Sno.	Metric	Description	Formula
1	Accuracy	It measures the correctly classified instances out of the total instances	TP+TNTP+TN+FP+FN
2	Specificity	It gives the proportion of true negatives to the amount of total negatives and False positives that the model predicts	TNTN+FP
3.	Recall/Sensitivity	Recall focuses on how good the model is at finding all the positives.	TPTP+FN
4	Confusion Matrix	A confusion matrix represents the predictive performance of a model on a dataset.	
5	Kappa	It is a statistical measure used to assess the agreement between predicted and actual classification in a classification problem.	k=Po-Pe1-Pe
6	Balanced accuracy	It is used to evaluate the performance of the classification model. it is the average of sensitivity and specificity	Sensitivity+Specificity2
7	The area under Receiver operating Characteristic curve (AUC)	AUC is the area under the ROC curve. It calculates the overall perfomance of a classification model across	0 means poor discriminative power of the model and 1 means excellent discriminative ability of the model

Source: Authors' compilation

5.4. Empirical analysis and discussion of the results

In this section, we conduct a comparative analysis of four machine learning models, logistic regression, and Random Forest, decision tree, and XGBOOST to cope with the loan default risk in Peer-to-Peer lending platforms.

Logistic Regression Model

The finds of our data analysis reveal that logistic regression analysis provide statistically significant results for all variables, except *"recoveries" and "dti,"* at the 5% and 1% significance levels. Therefore, on original dataset, the logistic regression model predicts loan defaults with an accuracy of 95%. However, it is noteworthy that the dataset has an imbalance, with just about 8%

of observations labelled as loan defaults and more than 92% as non-defaults. Given the imbalance in data, relying just on accuracy as a performance metric may be misleading. Moreover, our findings demonstrate that balanced accuracy outperforms accuracy while dealing with an imbalanced dataset. The balanced accuracy approach attempts to reduce classification error rates and should be preferable (Thölke et al., 2023). In a balanced dataset, accuracy and balanced accuracy tend to converge. As a result, model sensitivity and specificity, as well as balanced accuracy, must be prioritized. This approach ensures that the model effectively identifies both positive and negative instances, rather than merely concentrating on overall accuracy. By giving precedence to balanced accuracy, the model is better equipped to manage the dataset's imbalance and make more precise predictions regarding loan defaults. The model exhibits a sensitivity of 32.45% on the original dataset, demonstrating that it correctly predicts only one-third of actual defaults. The specificity is 99.99% but the balanced accuracy is 66.22%. Since the main focus is on reducing the Type II error and sensitivity is quite low and fails to identify the two-third of the loan default instances. Hence, this raises question on the performance of the model. The primary focus is on minimizing Type II error, the sensitivity is relatively low, indicating a failure to identify approximately two-thirds of loan default instances. This raises doubts regarding the model's performance and its effectiveness in accurately capturing loans at risk of default. In addition to this, to counter the imbalance problem, the minority oversampling technique is employed using the ROSE package in R, involving both under-sampling and oversampling. The implementation of synthetic sampling techniques, both under sampling and over-sampling resulted in a decrease in accuracy to 74.31% and 74.81% respectively. Especially, sensitivity improves to 68%, implying that out of 100 observations related to loan defaults, the model correctly predicts 68. The specificity of the model, which was firstly 99.99% in the original dataset, drops to 81.51% under synthetic sampling as there is always a trade-off between sensitivity and specificity. High specificity means the model's ability to correctly identify cases with no defaults, and the decline in specificity is related to the alterations made through over-sampling and under-sampling techniques.

Table 5: Performance Metrics of Logistic Regression Model

Parameter	Original dataset	Under-sampling (ROSE)	Over Sampling (ROSE)
Accuracy	95.13%	74.31%	74.81%
Sensitivity	32.45% *Lowest*	67.66%	68.11%

Specificity	99.99%	80.96%	81.51%
Balanced accuracy	66.22%	74.31%	74.81%
Kappa	0.4709	0.4862	0.4962
AUC	0.8311	0.8343	0.8358

Source: Authors' findings

Random Forest

The findings of the random forest model reveal that the accuracy is 95.44% as mentioned in Table 6, which means that out of 100 times, over 95 times it predicts the response outcome (0 or 1) correctly. The varImp in R is utilized to identify the strong and influential variables that affect the loan default prediction using Random Forest model. Based on mean decrease ginni, a metric to find the significant variable has identified recoveries made, total recovery of principal, total payments, interest rate, and dti as influential variables. As explained before the dataset is not balanced, therefore accuracy can be considered as an appropriate measure of performance metric. There is a need to look into the sensitivity value of the random forest model. The sensitivity value is 41.3 % which is low. Out of 100 cases of loan default, 41 can be identified correctly. The reason for the low value of sensitivity is the data imbalance in the original dataset. Table 6 also depict the results of random forest based on under-sampling and oversampling. There is an improvement in accuracy and sensitivity when the oversampling technique is applied. The accuracy was approximately 97% and the sensitivity is over 99% which means 99 times out of 100 times it correctly predicts the loan default (positive instances). The result of under-sampled dataset has lesser accuracy (77%) and reduced sensitivity compared to the oversampled dataset as illustrated in Table 6.

Table 6: Performance Metrics based on Random Forest Model

Parameter	**Original dataset**	**Under-Sampling (ROSE)**	**Over Sampling (ROSE)**
Accuracy	95.44%	75.21%	97.27%
Sensitivity	41.3%	72.88 %	99.4%
Specificity	99.64%	77.54%	95.13 %
Balanced accuracy	70.47%	75.21%	97.27%
Kappa	0.545	0.504	0.945
AUC	0.792	0.829	0.9959

Source: Authors' findings

The Cohen's kappa is 0.545 on the imbalanced dataset which depicts that there is moderate agreement between the predicted and actual classifications for loan default prediction. As far as the under-sampled dataset is concerned, the kappa value is 0.504 as it witnessed a slight decrease. However, on oversampling data, the kappa value has shown a magnificent increase in value. Cohen's kappa is 0.945 depicts almost perfect agreement between predicted and actual classifications. Adding further, the use of ROSE has positively impacted the model's ability to generalize the minority class.

Decision Tree Model

The findings explain that while the accuracy of the original dataset is remarkable, above 95%, caution is warranted due to the imbalanced nature of the dataset. Relying solely on accuracy can be deceptive in such cases. Other performance measures such as sensitivity and kappa become more important in assessing model performance. On the original dataset, the sensitivity is found to be 32.78%, showing that less than one-third of loan defaults are correctly predicted by the model. The kappa value stands at 0.475, suggesting a moderate level of agreement between predicted and actual classes. When applying the decision tree model to an under sampled balanced dataset, a decrease in accuracy is observed. However, the sensitivity remarkably improves to almost 78%, indicating that more than three-fourths of positive instances are correctly predicted. Despite, this improvement, the kappa value slightly decreases to 0.434, showing a reduction in the moderate level of agreement between predicted and actual classes. In contrast, when the decision tree model is applied to an oversampled dataset, the sensitivity witnesses a decline to 69.33%, compared to the 77.68% observed in the under-sampled dataset as illustrated by Table 7. Regardless of the decrease, the sensitivity more than doubles when compared to the original data. The specificity increases to 75.53%, indicating an improvement in the correct prediction of negative instances, but there is no significant improvement in Cohen's kappa.

Table 7: Performance Matrix based on Decision Tree Model

Parameter	Original dataset	Under-Sampling (ROSE)	Over Sampling (ROSE)
Accuracy	95.08%	71.70%	72.53%
Sensitivity	32.78%	77.68%	69.33%
Specificity	100%	65.72%	75.73%
Balanced accuracy	66.39%	71.70%	72.53%
Kappa	0.475	0.434	0.451
AUC	0.67	0.72	0.73

Source: Authors' findings

The results from the XGBOOST model highlight its significant accuracy, surpassing 95% on the original dataset. However, given the imbalance nature of the dataset, it is important to focus on sensitivity, which assesses the ability of the model to correctly identify instances of loan default. The sensitivity of approximately 40% implies that out of 100 instances of loan default, the model correctly predicts around 40, signifying a challenge in predicting defaults due to the higher occurrence of no default cases. Subsequently, the model tops in correctly identifying instances of no default, leading to an almost perfect specificity of 100%. The kappa value of 0.549 suggests a moderate level of agreement between the predicted and actual classes, highlighting the model's overall performance. The run of importance using *xgb.importance* function in R exhibited that Recoveries made, interest rate, total recovery of principal, installments paid, grade categories are the influential variables for loan default prediction.

When applied XGBOOST model is on an under-sampled dataset, a decline in accuracy is observed settling at 79.36%. The decline can be attributed to a reduction in the imbalances in the dataset. Despite the decline in the accuracy of the model, the sensitivity of the model reached around 78% nearly doubling compared to the sensitivity of the original dataset, however, the increase in sensitivity comes at a cost of specificity, which witnesses a decline to 81%, down from almost 100% in the original dataset. The kappa value also shows modest improvement.

In contrast, applying XGBOOST to an oversampled dataset results in a noteworthy decline in accuracy to 80.53% from the over 95% detected on the original dataset. However, this reduction in accuracy is supplemented by an improvement in sensitivity, while the specificity value experiences a decline. Particularly, the kappa value in this instance stands at 0.6107, surpassing the kappa value of the original dataset. This signifies a higher level of agreement between the predicted and actual classes in the oversampled dataset.

Table 8: Performance Matrix based on XGBOOST Model

Parameter	Original dataset	Under sampling (ROSE)	Over Sampling (ROSE)
Accuracy	95.56%	79.36%	80.53%
Sensitivity	39.53 %	77.73%	79.33%
Specificity	99.98%	80.98%	81.74%
Balanced accuracy	69.76%	79.36%	80.53%
Kappa	0.5469	0.5872	0.6107
AUC	0.8926	0.8893	0.901

Source: Authors findings

Table 9: Comparison between Logistic Regression, Random Forest Model, Decision Tree, and XGBOOST on original dataset

Model	Accuracy	Sensitivity	Specificity	Kappa	AUC
Logistic Regression	95.13%	32.45%	99.99%	0.4709	0.8311
Random Forest	95.44%	41.30%	99.64%	0.545	0.792
Decision Tree	95.08%	32.78%	100%	0.479	0.664
XGBOOST	95.56%	39.53 %	99.98	0.5469	0.8926

Source: Authors findings

In the evaluation of different machine learning models for loan default predictions, the Logistic Regression model exhibited an accuracy of 95.13%, indicating a high level of overall correct predictions. However, its sensitivity was relatively low at 32.45%, reflecting a limited ability to capture true positives. On the other hand, the specificity was exceptionally high at 99.99%, demonstrating accurate identification of true negatives. The Kappa coefficient suggested a moderate level of agreement (0.4709), and the Area under the Curve (AUC) value was 0.8311, indicating good discriminative ability. Comparatively, the Random Forest model showed a slightly higher accuracy at 95.44% than Logistic Regression. It demonstrated an improved sensitivity of 41.3%, signifying an enhanced ability to capture true positives. The specificity remained high at 99.64%, and the Kappa coefficient indicated a moderate to a good level of agreement (0.545). The AUC value for Random Forest was 0.792, reflecting good discriminative ability. The XGBOOST model has been the best model when applied on imbalanced dataset with an accuracy of 95.56% and AUC of 0.8926 whereas decision tree emerges out to be weaker model as the accuracy level and sensitivity is low compared to the other models.

Table 10: Comparison between Logistic Regression, Random Forest Model, Decision Tree, and XGBOOST on under-sampled dataset

Model	Accuracy	Sensitivity	Specificity	Kappa	AUC
Logistic Regression	74.31%	67.66%	80.96%	0.4862	0.8343
Random Forest	75.21%	72.88 %	77.54%	0.504	0.829
Decision Tree	71.70%	77.68%	65.72%	0.434	0.72
XGBOOST	79.36%	77.73%	80.98%	0.5872	0.8893

Source: Authors findings

In the evaluation of under sampled datasets, each machine learning model demonstrated varying degrees of performance for predicting loan defaults. Logistic Regression exhibited moderate

accuracy at 74.31%, with a sensitivity of 67.66%, indicating a moderate capability in capturing true positives. The specificity was 80.96%, signifying moderate accuracy in correctly identifying true negatives. The Kappa coefficient suggested a moderate level of agreement (0.4862), and the Area under the Curve (AUC) value was 0.8343, indicating good discriminative ability. Similarly, Random Forest displayed moderate accuracy at 75.21%, showcasing a good sensitivity of 72.88% for capturing true positives. Specificity was 77.54%, demonstrating moderate accuracy in correctly identifying true negatives. The Kappa coefficient indicated a moderate level of agreement (0.504), and, the AUC value was 0.829, reflecting good discriminative ability. The decision tree model accuracy level is lowest (71.70%), and AUC is also lowest but sensitivity level is relatively higher when compare with random forest model. Still decision tree model is a weak model. XGBOOST Model has shown a higher level of accuracy, sensitivity, specificity, and AUC which makes it the best model for undersampled balanced dataset.

Table 11: Comparison between Logistic Regression, Random Forest Model, Decision Tree, and XGBOOST on Over-sampled dataset

Model	Accuracy	Sensitivity	Specificity	Kappa	AUC
Logistic Regression	74.81%	68.11%	81.51%	0.4962	0.8358
Random Forest	97.27%	99.4%	95.13%	0.945	0.9959
Decision Tree	72.53%	69.33%	75.73%	0.451	0.73
XGBOOST	80.53%	79.33%	81.74%	0.6107	0.901

Source: Authors' findings

In the context of the oversampled dataset, various machine learning models were assessed for their performance in predicting loan defaults. Logistic Regression demonstrated a moderate accuracy of 74.81%, with good sensitivity at 68.11% for effectively capturing true positives. The specificity stood at 81.51%, indicating moderate accuracy in correctly identifying true negatives. The Kappa coefficient suggested a moderate level of agreement (0.4962), and the Area under the Curve (AUC) value was 0.8358, signifying good discriminative ability. Random Forest exhibited notably high accuracy at 97.27%, with excellent sensitivity (99.4%) for capturing true positives. The specificity was also high at 95.13%, demonstrating accuracy in correctly identifying true negatives. The Kappa coefficient indicated a very high level of agreement (0.945), and the AUC value was an outstanding 0.9959, showcasing excellent discriminative ability. The decision tree model has accuracy of 72.53%, sensitivity of 69.33%, and AUC of 0.73 which is relatively lower when compare to other models. The XGBOOST model shows a greater accuracy of 80.53 but still lower

compared to random forest model. The sensitivity and AUC of XGBOOST model is 79.33% and 0.901 but still lags behind the random forest model. XGBOOST emerges as a runner-up model, whereas least preferred models are decision tree and logistic regression.

6. Conclusion and Implications of the Study

Given that borrowers predominantly belong to the low-income population bracket, our findings are particularly applicable to developing or underdeveloped nations. The study concludes that the XGBOOST model performs better than the logistic regression model, Random Forest, and Decision tree in predicting loan default on both the original dataset and the balanced dataset (Under sampled balanced dataset).

However, when the analysis is performed on oversampled balanced dataset, the Random forest model outperforms all the other models. Although the performance of XGBOOST is satisfactory but it still falls short. The logistic regression and decision tree model have not given satisfactory result. Thus, the study concludes that the Random Forest model is superior when applied to over sampled balanced datasets and XGBOOST emerges as best model for under sampled balanced dataset and original dataset. The dataset problem could be solved with accumulation of actual data, hence for future references in emerging economies XGBOOST model can help in minimizing NPA's for FinTechs and banks. On oversampled data, Random forest model is better equipped to correctly identify the true positives and negatives with high specificity, accuracy and sensitivity.

The borrower's default risk poses a major threat to lending platforms and financial institutions. A high default risk leads to sizeable non-performing assets (NPAs), resulting in significant losses for these platforms. This study is valuable for lending platforms and banks as it proposes models that can predict borrower defaults with greater accuracy. Employing the appropriate model will enhance the overall performance of default prediction, probably saving these platforms millions of dollars that might otherwise be lost due to borrower misclassification. Even a slight improvement in predictive performance using machine learning models can significantly impact the profitability of these platforms and will lead to better risk management.

7. Limitation of the study and Scope for Future Research

Every research work inherently encounters limitations. The current study is no exception. Firstly, the logistic model assumes linearity and may not perform well if the relationship is non-linear. Secondly, the random forest model suffers from the overfitting problem as the train dataset

captures the noises and random fluctuation so it may not perform well on the test data. The n tree for the random forest is kept at 5 to reduce the overfitting problem. Thirdly, the XGBOOST model, though depicts good predictive performance but it is prone to overfitting and sensitive to hyper parameter tuning. Lastly, decision tree model is majorly bias toward majority class which reduces the predictive accuracy compared to ensemble methods such as RF, XGBOOST, and, gradient boosting.

A future study could examine the effectiveness of combining multiple models through ensemble methods. Adding further, there is prospect to introduce temporal analysis, examining the performance of loan default prediction models over varying time periods, economic conditions, and regulatory environments. This includes analyzing changes in model performance over time.

References

Amstad, M. (2019). Regulating Fintech: Objectives, principles, and practices. *Asian Development Bank Institute Working Paper Series*, *1016*.

Anil, K., & Misra, A. (2022). Artificial intelligence in Peer-to-peer lending in India: A cross-case analysis. *International Journal of Emerging Markets*, *17*(4), 1085–1106.

Appio, F. P., La Torre, D., Lazzeri, F., Masri, H., & Schiavone, F. (n.d.). Artificial Intelligence in Business: Opportunities and Challenges. *Impact of Artificial Intelligence in Business and Society*, 83–167.

Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, *14*(1), 106. https://doi.org/10.1186/1471-2105-14-106

Buchak, G., Matvos, G., Piskorski, T., & Seru, A. (2018). Fintech, regulatory arbitrage, and the rise of shadow banks. *Journal of Financial Economics*, *130*(3), 453–483.

Chang, Y.-C., Chang, K.-H., Chu, H.-H., & Tong, L.-I. (2016). Establishing decision tree-based short-term default credit risk assessment models. *Communications in Statistics - Theory and Methods*, *45*(23), 6803–6815. https://doi.org/10.1080/03610926.2014.968730

Coşer, A., Maer-Matei, M. M., & Albu, C. (2019). PREDICTIVE MODELS FOR LOAN DEFAULT RISK ASSESSMENT. *Economic Computation & Economic Cybernetics Studies & Research*, *53*(2). https://ecocyb.ase.ro/nr2019_2/9.%20Coser%20Al.%20Crisan%20Albu%20(T).pdf

Dansana, D., Patro, S. G. K., Mishra, B. K., Prasad, V., Razak, A., & Wodajo, A. W. (2024). Analyzing the impact of loan features on bank loan prediction using R andom F orest algorithm. *Engineering Reports*, *6*(2), e12707. https://doi.org/10.1002/eng2.12707

Emekter, R., Tu, Y., Jirasakuldech, B., & Lu, M. (2015). Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. *Applied Economics*, *47*(1), 54–70. https://doi.org/10.1080/00036846.2014.962222

Ferozi, M. J. (2018). *Loan data for Lending club* [Dataset]. Kaggle.

Granström, D., & Abrahamsson, J. (2019). *Loan default prediction using supervised machine learning algorithms*.

Guo, Y. (2020). Credit risk assessment of P2P lending platform towards big data based on BP neural network. *Journal of Visual Communication and Image Representation*, *71*, 102730.

Hegde, H. A., & Bhowmik, B. (2023, December). Feature Selection for Peer-to-Peer Lending Default Risk Using Boruta and mRMR Approach. In 2023 IEEE 20th India Council International Conference (INDICON) (pp. 983-988). IEEE.

Jiang, Z., Ma, G., & Zhu, W. (2022). Research on the impact of digital finance on the innovation performance of enterprises. *European Journal of Innovation Management*, *25*(6), 804–820.

Johnson, P. C., Laurell, C., Ots, M., & Sandström, C. (2022). Digital innovation and the effects of artificial intelligence on firms' research and development–Automation or augmentation, exploration or exploitation? *Technological Forecasting and Social Change*, *179*, 121636.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.

Kohv, K., & Lukason, O. (2021). What best predicts corporate bank loan defaults? An analysis of three different variable domains. *Risks*, *9*(2), 29.
Lai, L. (2020). Loan default prediction with machine learning techniques. *2020 International Conference on Computer Communication and Network Security (CCNS)*, 5–9. https://ieeexplore.ieee.org/abstract/document/9240729/
Li, Z., Li, S., Li, Z., Hu, Y., & Gao, H. (2021). *Application of XGBoost in P2P default prediction*. *1871*(1), 012115.
Liu, Y., Yang, M., Wang, Y., Li, Y., Xiong, T., & Li, A. (2022). Applying machine learning algorithms to predict default probability in the online credit market: Evidence from China. *International Review of Financial Analysis*, *79*, 101971.
Lubis, R. M. F., & Huang, J. P. (2024). Leveraging Machine Learning to Predict Credit Card Customer Segmentation. Journal of Ecohumanism, 3(7), 3386-3418.
Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., & Niu, X. (2018). Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. *Electronic Commerce Research and Applications*, *31*, 24–39.
Madaan, M., Kumar, A., Keshri, C., Jain, R., & Nagrath, P. (2021). Loan default prediction using decision trees and random forest: A comparative study. *IOP Conference Series: Materials Science and Engineering*, *1022*(1), 012042. https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012042/meta
Moula, F. E., Guotai, C., & Abedin, M. Z. (2017). Credit default prediction modeling: An application of support vector machine. *Risk Management*, *19*(2), 158–187. https://doi.org/10.1057/s41283-017-0016-x
Muslim, M. A., Nikmah, T. L., Pertiwi, D. A. A., & Dasril, Y. (2023). New model combination meta-learner to improve accuracy prediction P2P lending with stacking ensemble learning. *Intelligent Systems with Applications*, *18*, 200204.
Nalić, J., & Švraka, A. (2018). *Using data mining approaches to build credit scoring model: Case study—Implementation of credit scoring model in microfinance institution*. 1–5.
Rymarczyk, T., Kozłowski, E., Kłosowski, G., & Niderla, K. (2019). Logistic regression for machine learning in process tomography. *Sensors*, *19*(15), 3400.
Ryu, H., & Chang, Y. (2018). What makes users willing or hesitant to use Fintech. *The Moderating Effect of User Type*.
Sam'an, M., Safuan, S., & Munsarif, M. (2024). Feature selection in P2P lending for default prediction using grey wolf optimization and machine learning. Bulletin of Electrical Engineering and Informatics, 13(5), 3609-3615.
Syed Nor, S. H., Ismail, S., & Yap, B. W. (2019). Personal bankruptcy prediction using decision tree model. *Journal of Economics, Finance and Administrative Science*, *24*(47), 157–170.
Thölke, P., Mantilla-Ramos, Y.-J., Abdelhedi, H., Maschke, C., Dehgan, A., Harel, Y., Kemtur, A., Berrada, L. M., Sahraoui, M., & Young, T. (2023). Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. *NeuroImage*, *277*, 120253.
Uddin, N., Ahamed, M. K. U., Uddin, M. A., Islam, M. M., Talukder, M. A., & Aryal, S. (2023). An ensemble machine learning based bank loan approval predictions system with a smart application. *International Journal of Cognitive Computing in Engineering*, *4*, 327–339.
Vallee, B., & Zeng, Y. (2019). Marketplace lending: A new banking paradigm? *The Review of Financial Studies*, *32*(5), 1939–1982.
Vives, X. (2020). Digital disruption in banking and its impact on competition. *Organisation for Economic Co-Operation and Development (OECD)*.
Wang, J. (2022). Prediction of postoperative recovery in patients with acoustic neuroma using machine learning and SMOTE-ENN techniques. *Math. Biosci. Eng*, *19*(10), 10407–10423.
Xu, J. J. (2016). Are blockchains immune to all malicious attacks? *Financial Innovation*, *2*, 1–9.
Xu, J., Lu, Z., & Xie, Y. (2021). Loan default prediction of Chinese P2P market: A machine learning methodology. *Scientific Reports*, *11*(1), 18759.
Zhou, J., Li, W., Wang, J., Ding, S., & Xia, C. (2019). Default prediction in P2P lending from high-dimensional data based on machine learning. *Physica A: Statistical Mechanics and Its Applications*, *534*, 122370.
Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. *Procedia Computer Science*, *162*, 503–513.
Zhu, X., Chu, Q., Song, X., Hu, P., & Peng, L. (2023). Explainable prediction of loan default based on machine learning models. *Data Science and Management*.