



Comparative Analysis of Loan Prediction Models with Imbalanced Data and Impact of Loan Eligibility Metrics

Jeremiah Philip

Applied Research Project submitted in partial fulfilment of the requirements for
the degree of MSc Financial Analytics
at Dublin Business School

Supervisor: Dr. Monika Sosa Smatralova

August 2023

Declaration

I, Jeremiah Philip, declare that this Applied Research Project that I have submitted to Dublin Business School for the award of MSc Financial Analytics is the result of my own investigations, except where otherwise stated, where it is clearly acknowledged by references. Furthermore, this work has not been submitted for any other degree.

Signed: Jeremiah Philip

Student Number: 10623819

Date: 28-08-2023

Acknowledgment

I would like to express my heartfelt gratitude to Dublin Business School for their support in making this research possible. I am deeply thankful to God for granting me the opportunity and mental strength to undertake and complete this journey. I am also grateful to my family for their unwavering financial and emotional support. My sincere appreciation goes to my friends for their valuable assistance during this project. I would also like to thank my supervisor Dr. Monika Sosa Smatralova for her support throughout the phase of this project.

Table of Contents

<i>Declaration</i>	2
<i>Acknowledgment</i>	3
<i>List of Tables</i>	5
<i>List of Figures</i>	5
<i>Abstract</i>	6
<i>Introduction</i>	7
Problem Statement.....	7
Motivation and Challenges	8
Research Objectives.....	9
Research Question	10
<i>Literature Review</i>	12
<i>Methodology</i>	22
Business Understanding.....	23
Data Understanding	23
Analysis of Data	25
Data Preparation	27
Loan Eligibility Metrics.....	28
Data Imbalance Handling.....	29
Model Training	31
Model Evaluation	31
Comparison of Models.....	32
<i>Results And Discussion</i>	34
Impact of Loan Eligibility Metrics on Loan Default Prediction Models	35
Performance of Resampling Techniques	35
Comparison of Decision Tree and Random Forest Models.....	36
Validation of Results.....	36
Critical Evaluation of Previous Research Results.....	37
<i>Conclusion and Future works</i>	41
<i>References</i>	43
<i>Appendices</i>	45

List of Tables

<i>Table 1: Summary of Related Works</i>	20
<i>Table 2: Results of Decision Tree Model</i>	34
<i>Table 3: Results of Random Forest Model</i>	35
<i>Table 4: Validation of Results</i>	36
<i>Table 5: Performance Comparison of Related Works with Proposed System</i>	38

List of Figures

<i>Figure 1: Research workflow and Methodology</i>	22
<i>Figure 2: Loan Status Distribution</i>	25
<i>Figure 3: Correlation Matrix</i>	26
<i>Figure 4: Distribution of Key Attributes Dataset</i>	45
<i>Figure 5: Scatter Plot - Person Income vs Loan Amount</i>	45
<i>Figure 6 : Data Visualization - Bar Plots of Categorical Features</i>	46
<i>Figure 7: Decision Tree Results Visualization</i>	46
<i>Figure 8: Random Forest Results Visualization</i>	46

Abstract

Loan prediction models play a vital role in determining borrowers' likelihood of defaulting on loans, but their development is challenging when dealing with imbalanced datasets. This research investigated the impact of including loan eligibility metrics on the performance of balanced loan default prediction models. Two machine learning models, Decision Tree and Random Forest, were compared in handling imbalanced data. To address data imbalance, Synthetic Minority Oversampling Technique (SMOTE), Under Sampling, and Random Over Sampling were used. The study validates the proposed methodology using a dataset from Kaggle. The findings revealed that incorporating loan eligibility metrics significantly improves the accuracy of balanced loan default prediction models. Among the models, Random Forest stands out, achieving the highest accuracy of 93.67%. This research contributes to financial analytics and data science, offering an optimized loan prediction model that empowers banks to enhance their loan decision-making process and effectively manage credit risk.

Key Words: Loan Prediction, Imbalanced Data, Loan Eligibility Metrics, Decision Tree, Random Forest, Synthetic Minority Oversampling Technique (SMOTE), Random Over Sampling, Under Sampling.

Introduction

Loan prediction models played an essential role in the banking industry, enabling financial institutions to assess the likelihood of borrowers defaulting on loans. These models aided in the decision-making process by evaluating the creditworthiness of applicants and predicting their ability to repay loans. However, developing accurate loan prediction models proved to be challenging, especially when dealing with imbalanced datasets where instances of default were relatively rare compared to non-default cases. This research aimed to address this challenge by comparing the performance of two machine learning models, Decision Tree, and Random Forest, in handling imbalanced datasets. Additionally, the study aimed to evaluate the impact of including loan eligibility metrics in the dataset on model accuracy. By examining the effects of loan eligibility metrics and exploring techniques to handle imbalanced data, this research sought to contribute to the field of financial analytics and data science.

Problem Statement

In the context of the banking industry, the evaluation of loan applications were essential for determining the creditworthiness of borrowers and predicting the likelihood of loan default. Predictive models based on machine learning techniques have been developed to enhance the loan decision-making process. However, the presence of imbalanced data in loan prediction poses a significant challenge, leading to biased models that perform poorly in predicting loan defaults. Furthermore, the impact of inclusion of loan eligibility metrics in the dataset has not been thoroughly explored in relation to improving loan default classification. This research aimed to address the problem of imbalanced data in loan prediction models and investigate the impact of including loan eligibility metrics on the performance of loan default prediction. By comparing the effectiveness of Decision Tree and Random Forest algorithms in handling imbalanced datasets, this study sought to identify the most suitable machine learning algorithm for loan default prediction. Additionally, the study aimed to evaluate the performance of

resampling techniques, including SMOTE, Under Sampling, and Random Over Sampling Technique, in mitigating the imbalance issue. The original contribution of this research lies in the comprehensive investigation of loan eligibility metrics and imbalanced data handling techniques in loan default classification. By providing insights into the significance of loan eligibility metrics and the effectiveness of different algorithms and resampling techniques, this study aimed to contribute to the advancement of financial analytics and data science. To achieve these research objectives, a dataset collected from Kaggle was utilized, and a systematic methodology was followed, including data preprocessing, model training, model evaluation using various performance metrics, and comparative analysis of the models with and without loan eligibility metrics. The findings from this research provided valuable insights into building an optimized loan prediction model that enhances the loan decision-making process for banks, leading to improved credit risk management and more informed loan decisions.

Motivation and Challenges

The motivation behind this research derives from the necessity of streamlining and enhancing the loan application process within the banking sector. With a personal experience of the loan application process, the researcher was driven to explore avenues for improving its efficiency. Traditional loan evaluations impose time and complexity burdens on both lenders and borrowers. By developing more accurate and efficient loan prediction models, the research endeavors to elevate the efficiency of the loan application process and extend access to credit facilities for those in need. The primary challenge addressed in this research was the issue of imbalanced datasets. In loan prediction tasks, instances of loan defaults are relatively rare compared to non-default cases. This imbalance poses a significant obstacle to developing robust loan prediction models that can accurately identify default cases. Conventional machine learning algorithms often struggle to effectively classify rare events due to their bias towards

the majority class. Therefore, this research aimed to compare the performance of Decision Tree and Random Forest models in handling imbalanced datasets and provide insights into improving loan default classification.

Another challenge arised from the inclusion of loan eligibility metrics in the dataset. Loan eligibility metrics, such as debt-to-income ratio, loan-to-income ratio, and employment length ratio, introduce complex interactions that impact the accuracy of loan default prediction models. Understanding the influence of these metrics on model performance were crucial for developing reliable loan prediction models. Justifying the significance of addressing these challenges, accurate loan default prediction are crucial for financial institutions to effectively manage credit risks. By accurately identifying borrowers likely to default, banks can make informed decisions regarding loan approvals, loan terms, and risk mitigation strategies. Moreover, improving loan default classification can streamline the loan application process, reducing the time taken to evaluate applications and increasing access to credit facilities for deserving individuals.

Research Objectives

Drawing from the insights presented in the subsequent literature review, the primary research objective were highlighted as follows:

1. Examine whether the inclusion of loan eligibility metrics, such as debt-to-income ratio, loan-to-income ratio, and employment length ratio, results in enhanced predictive accuracy of loan default models.

Additionally, the following objective was formulated:

2. Compare the performance of Decision Tree and Random Forest algorithms in handling imbalanced datasets for loan default prediction, and investigate the impact of various

resampling techniques, such as Synthetic Minority Oversampling Technique (SMOTE), Under Sampling, and Random Over Sampling, on the accuracy of loan prediction models. By accomplishing these research objectives, this study aimed to contribute to the field of financial analytics and data science by providing a comprehensive understanding of the performance of machine learning algorithms in handling imbalanced datasets for loan default prediction. Moreover, the research sought to elucidate the impact of loan eligibility metrics on loan default predictions and offer practical recommendations for the development of optimized loan prediction models that can benefit financial institutions in their loan decision-making processes.

Research Question

Loan prediction models are essential for banks to assess the likelihood of borrowers defaulting on loans. However, the issue of imbalanced data in loan prediction poses a significant challenge for the application of machine learning algorithms. In real-world scenarios, datasets are often imbalanced, with the majority class outweighing the minority class. This imbalance can lead to models that are biased towards the majority class, resulting in poor performance in predicting the minority class. Therefore, this researcher have explored various techniques for handling imbalanced data, including resampling methods.

The problem of imbalanced data are prevalent in loan prediction, and it affects the accuracy of loan default prediction, which were a critical aspect of credit risk management. Solving this problem, along with the inclusion of loan eligibility metrics in the dataset, are crucial for the financial health of banks and other financial institutions. Accurate loan default prediction helps banks to avoid credit losses, maintain positive cash flow, and make informed loan decisions. Moreover, solving the problem of imbalanced data in loan prediction enhances the effectiveness of machine learning algorithms, which are a powerful tool for predicting loan

defaults. The study aimed to contribute to the field of financial analytics and data science by investigating the effects of loan eligibility metrics and imbalanced data handling techniques on loan default classification. Specifically, the research questions addressed include:

Primary Question

- (1) What is the impact of including loan eligibility metrics in the dataset on the performance of balanced loan default prediction models?

In addition questions

- (2) How do resampling techniques, such as SMOTE, Under sampling and Random over sampling impact the performance of loan default prediction models?
- (3) Which machine learning algorithm, Decision Tree or Random Forest, is more effective in handling imbalanced datasets for loan default prediction?

Literature Review

A banking institution are a financial organization that engages in the receipt and transfer of deposits, as well as lending activities. These lending activities include the provision of loans, advances, and other credit facilities to the public, using capital from deposited funds [1]. The role of banks in a market economy are significant. In assessing loan applications, banks attempt to evaluate the creditworthiness of borrowers, categorizing them as either good (non-defaulter) or bad (defaulter). The ability to predict the likelihood of loan default is crucial to the lender [2]. Loan default occurs when a borrower fails to repay a loan as per the agreed-upon terms. Default prediction involves the use of historical and current data, credit and payment information, and customer credit behavior to forecast a borrower's capacity to repay a loan while assessing loan profitability. Numerous machine learning classification models have been used for this purpose [1]. However, despite the availability of these models, the loan application process can still be time-consuming, with a high volume of applicants seeking loans for various purposes such as house loans and car loans. As someone who has personally experienced this process, researcher was motivated to explore ways to improve the efficiency of the loan application process. By developing more accurate and efficient machine learning models to predict loan default, the time taken to evaluate loan applications can be reduced and make the lending process more accessible to individuals seeking credit facilities.

In the banking industry, loan prediction models play a critical role in assessing the creditworthiness of potential borrowers. By analyzing various factors, including loan eligibility metrics, these models can help banks make informed decisions on whether to grant a loan or not. However, the development of an accurate loan prediction model was not without its challenges, especially when dealing with imbalanced datasets where the occurrence of defaults are rare. This research aimed to address this issue by comparing the performance of two machine learning models, Decision Tree and Random Forest, in handling imbalanced data and

evaluating the impact of loan eligibility metrics on model accuracy. To overcome the issue of imbalanced data, Synthetic Minority Oversampling Technique (SMOTE), Under sampling and Random Over Sampling Technique was utilized. Moreover, including loan eligibility metrics in the dataset can have a substantial impact on loan default prediction accuracy. Loan eligibility metrics such as debt-to-income ratio, loan-to-income ratio and employment length ratio provide valuable information about a borrower's financial health and creditworthiness, which can significantly affect loan default prediction accuracy. Therefore, there was a need to explore and evaluate various methodologies to address the issue of imbalanced data in loan prediction, as well as the impact of loan eligibility metrics on loan default prediction accuracy, to develop accurate models that can help financial institutions make informed loan decisions.

Machine learning algorithms leverage anonymized historical data to develop new models that enhance predictive accuracy. A well-designed model helps financial institutions anticipate the probability of loan default and implement necessary precautions before default occurs [3]. There are no one-size-fits-all machine learning model that suitable for all situations. Instead, each machine learning system consists of several components, including the problem statement, data source, model architecture, optimization algorithm, and validation and testing frameworks [4]. In real-world scenarios, imbalanced datasets are prevalent, such as those encountered in fraud detection, risk management, and medical diagnosis. Predicting outcomes for such datasets poses a significant challenge as classifiers tend to favor the majority class at the expense of the minority class.

Several studies have been conducted to address the issue of imbalanced data in loan prediction. The research conducted by [1] aimed to address the issue of loan default prediction by comparing the performance of three classification methods: Naïve Bayes, Decision Tree, and Random Forest. The study recognized the importance of data engineering, including data

preprocessing techniques and feature selection, in improving the accuracy of the prediction models. The researchers applied various pre-processing techniques to handle challenges such as missing values and imbalanced data, which are common issues in loan prediction. The results of the study demonstrated the significant impact of data engineering on the performance of the models. The best-performing model achieved an improvement of approximately 40% in prediction accuracy, highlighting the effectiveness of the applied data preprocessing techniques and feature selection. This finding emphasizes the importance of careful data preparation in machine learning tasks, as it can greatly enhance the accuracy and performance of the models.

However, it is important to note that the study did not include a comparative analysis of imbalanced data on the prediction accuracy of different models. Imbalanced data is a common challenge in loan prediction, where the occurrence of loan defaults is relatively rare compared to non-default cases. This imbalance can lead to biased predictions and lower accuracy, particularly for the minority class. Considering the impact of imbalanced data on the prediction accuracy of different models would have provided valuable insights into the suitability of each model for addressing this challenge. Additionally, the study did not explore the impact of loan eligibility metrics on loan default prediction. Loan eligibility metrics, such as debt-to-income ratio, loan-to-income ratio, and employment length ratio, provide important information about a borrower's financial health and creditworthiness. These metrics can significantly affect the prediction accuracy of loan defaults. Evaluating the influence of loan eligibility metrics on the performance of the models would have contributed to a more comprehensive understanding of the factors influencing loan default prediction accuracy. In summary, while the research conducted by [1] made significant contributions by highlighting the importance of data engineering techniques in improving the accuracy of loan default prediction models, there are some aspects that were not addressed. A comparative study of imbalanced data on the prediction accuracy of different models and the evaluation of loan eligibility metrics' impact

on loan default prediction would have added further insights to the research. Considering these aspects in future studies would contribute to a more comprehensive understanding of loan prediction models in the context of the banking industry.

Another study by [2] focused on Peer-to-Peer (P2P) lending, where lenders rely solely on borrower-provided information to assess default risk. The study identified the information asymmetry problem inherent in P2P lending, leading to imbalanced datasets with unequal proportions of fully paid and default loans. To address this issue, the study proposed a scheme that incorporated several machine learning models, re-sampling techniques, and cost-sensitive mechanisms to handle the imbalanced data. A key aspect of the study was the utilization of different resampling techniques to address the imbalanced nature of the P2P lending dataset. Among the techniques used, the Synthetic Minority Over-Sampling Technique (SMOTE) was employed to generate synthetic instances of the minority class, thereby increasing its representation. This approach aimed to improve the prediction accuracy for default risk by providing the models with a more balanced dataset. Additionally, the study applied cost-sensitive learning, which involved adjusting the cost function to account for the imbalanced data. By assigning a scalar α in the cost function, the loss from the minority class (default loans) was given more weight, ensuring that the models prioritize correctly classifying instances of default. The results of the study demonstrated that the proposed scheme effectively raised the prediction accuracy for default risk.

However, it is important to note the difference between the current research being evaluated and the previous study. In the current research, a more comprehensive approach was adopted by comparing multiple resampling techniques, including SMOTE, Under Sampling, and Random Over Sampling. By considering a range of techniques, the aim was to identify the most suitable approach for the specific dataset being analyzed. Furthermore, an additional

aspect considered in the current research is the evaluation of the impact of loan eligibility metrics on the performance of the loan prediction models. Loan eligibility metrics, such as debt-to-income ratio, loan-to-income ratio, and employment length ratio, provide valuable insights into a borrower's financial health and creditworthiness. By incorporating these metrics into the analysis, the aim was to determine their influence on the prediction accuracy of loan defaults. In summary, while the study conducted by [2] focused on P2P lending and proposed a scheme involving machine learning models, re-sampling techniques, and cost-sensitive mechanisms to address imbalanced data, the current research takes a more comprehensive approach by comparing and evaluating multiple resampling techniques and considering the impact of loan eligibility metrics. By conducting such a comparative analysis, the current research aims to provide a more thorough evaluation of the loan prediction models and their ability to accurately assess default risk in the context of the banking industry.

Another Study [3] an empirical comparison was made between different combinations of classifiers and resampling techniques to address the issue of imbalanced data in credit risk prediction for social lending markets. The study introduced a novel risk assessment methodology that considered the imbalanced nature of the data and evaluated the credit predictions using a G-mean measure, which aimed to mitigate bias towards the majority class. The study explored various resampling approaches to tackle the imbalance problem, categorizing them into three main categories: under-sampling, over-sampling, and hybrid methods. Under-sampling techniques such as random under-sampling (RUS) and instance hardness threshold (IHT) were employed to reduce the majority class instances. Over-sampling techniques, including random over-sampling (ROS), synthetic minority over-sampling technique (SMOTE), and adaptive synthetic sampling (ADASYN), were utilized to increase the minority class instances. Additionally, hybrid approaches such as SMOTE + Tomek links (SMOTE-TOMEK) and SMOTE + edited nearest neighbor (SMOTE-ENN) were considered

in the study. The results of the research indicated that combining the random forest classifier with random under-sampling yielded promising outcomes in calculating the credit risk associated with loan applicants in social lending markets. This combination demonstrated effectiveness in addressing the class imbalance problem and improving the accuracy of credit risk prediction.

However, it is important to note the differences between the aforementioned study and the current research being evaluated. In the current research, a comprehensive analysis was conducted by comparing multiple resampling techniques, including SMOTE sampling, under-sampling, and random over-sampling. Furthermore, the impact of loan eligibility metrics on the performance of the prediction models was evaluated. These metrics, such as debt-to-income ratio, loan-to-income ratio, and employment length ratio, provide valuable insights into a borrower's financial health and creditworthiness. Incorporating these metrics into the analysis aims to determine their influence on the prediction accuracy of loan defaults. By conducting such a comprehensive evaluation, the current research seeks to contribute to the existing body of knowledge by providing a deeper understanding of the effectiveness of different resampling techniques and the impact of loan eligibility metrics on credit risk prediction.

The authors [4] investigated credit risk prediction in the context of peer-to-peer (P2P) lending, where information asymmetry poses a challenge to accurately estimate default risk. The authors proposed a novel risk assessment methodology that incorporates imbalanced data and compared different combinations of classifiers and resampling techniques. Their findings emphasized the effectiveness of combining random forest and random under-sampling for predicting credit risk in social lending markets. While this study addressed the imbalanced data problem, it did not consider the comparative analysis of multiple resampling techniques or evaluate the impact of loan eligibility metrics. In contrast, the proposed research aims to

comprehensively compare the performance of Decision Tree and Random Forest algorithms in handling imbalanced datasets for loan default prediction. The inclusion of multiple resampling techniques, including Synthetic Minority Oversampling Technique (SMOTE), Under Sampling, and Random Over Sampling, allows for a thorough evaluation of their effectiveness. By considering these techniques, the research provides a more comprehensive analysis of addressing the imbalance issue. Additionally, the study evaluates the influence of loan eligibility metrics, such as debt-to-income ratio, loan-to-income ratio, and employment length ratio, on the accuracy of loan default prediction. This consideration of loan eligibility metrics further enhances the research's contribution to the field.

While the previous study focused on the combination of classifiers and resampling techniques, the proposed research takes a broader approach by comparing the performance of specific algorithms, employing multiple resampling techniques, and investigating the impact of loan eligibility metrics. This comprehensive analysis offers a more in-depth understanding of the factors influencing loan default prediction and provides valuable insights for financial institutions in enhancing their loan decision-making process. However, it is worth noting that both studies share the objective of addressing the challenge of imbalanced data in loan default prediction. The proposed research builds upon the findings of previous studies by expanding the scope of analysis, incorporating additional techniques, and considering the impact of loan eligibility metrics. By doing so, it aims to contribute to the existing knowledge in financial analytics and data science. Researcher (4) studies have made significant contributions to the field of loan default prediction by addressing the challenges associated with imbalanced datasets. The proposed research further advances this area by conducting a comprehensive analysis of algorithm performance, employing multiple resampling techniques, and evaluating the influence of loan eligibility metrics. The findings of this study are expected to enhance the

understanding of loan default prediction and provide practical insights for financial institutions in improving their loan decision-making processes.

Moving beyond the in-depth exploration of the four main literature reviews, it's important to recognize a wider range of relevant studies. The summarized collection, as shown in Table 1, covers different time periods and datasets, reflecting the ongoing effort to improve the accuracy of loan default prediction models. Each of these studies took unique paths, often using machine learning methods, strategies to address skewed data distribution, and considerations of loan eligibility factors. Some studies particularly dealt with methods like under-sampling, over-sampling, and SMOTE to handle the complexities arising from imbalanced data. Moreover, a few studies delved into incorporating cost-sensitive learning frameworks to fine-tune the prediction abilities of models. Of notable significance was the comprehensive analysis given to the four main literature reviews that align closely with the main goals of the ongoing research. These references closely resonate with the central themes of the current research, as evident in its goals and chosen methodologies. The current research similarly follows a parallel path, involving a thorough comparison of different machine learning approaches and the use of various resampling techniques like SMOTE, Under Sampling, and Random Over Sampling. This exploration also extends to investigating the role of loan eligibility metrics. These components, both on their own and in combination, play a pivotal role in shaping the research's framework.

Table 1: Summary of Related Works

Ref.	Year	Dataset Count	Machine Learning	Under Sampling	Over Sampling	SMOTE	Cost Sensitive	Loan Eligibility Metrics
[1]	2020	1	✓	✗	✓	✓	✗	✗
[2]	2021	1	✓	✓	✗	✗	✓	✗
[3]	2018	1	✓	✓	✓	✓	✗	✗
[4]	2020	1	✓	✗	✓	✓	✗	✗
[5]	2020	3	✓	✓	✓	✓	✗	✗
[6]	2012	1	✓	✗	✓	✗	✓	✗
[7]	2016	1	✓	✓	✗	✓	✗	✗
[8]	2018	1	✓	✓	✓	✓	✗	✗
[9]	2019	1	✓	✗	✓	✓	✗	✗
My Study	2023	1	✓	✓	✓	✓	✗	✓

The reviewed literature on loan default prediction has provided valuable insights into the challenges associated with imbalanced data and the effectiveness of different approaches. Several critical evaluations can be made regarding the previous research studies. Firstly, [1] focused on comparing the performance of three classification methods and emphasized the importance of data engineering techniques in improving prediction accuracy. However, the study did not include a comparative analysis of imbalanced data on the prediction accuracy of different models or evaluate the impact of loan eligibility metrics, which are crucial factors in loan default prediction. Secondly, [2] addressed the issue of imbalanced data in P2P lending by proposing a scheme that incorporated machine learning models, re-sampling techniques, and cost-sensitive mechanisms. While the study demonstrated the effectiveness of the proposed scheme, it did not compare multiple resampling techniques or evaluate the impact of loan eligibility metrics. Additionally, [3] conducted an empirical comparison of classifiers and resampling techniques in credit risk prediction for social lending markets. The study highlighted the importance of considering imbalanced data and used a G-mean measure to

mitigate bias. However, the study did not evaluate the impact of loan eligibility metrics or compare the performance of different resampling techniques comprehensively. Finally, [4] investigated credit risk prediction in P2P lending and proposed a risk assessment methodology that considered imbalanced data. Although the study achieved promising results by combining random forest and random under-sampling, it did not compare multiple resampling techniques or consider the influence of loan eligibility metrics. In contrast, the current research aims to address the limitations of previous studies by conducting a comprehensive evaluation. It compares the performance of Decision Tree and Random Forest models, employs multiple resampling techniques (SMOTE sampling, Under Sampling, and Random Over Sampling), and evaluates the impact of loan eligibility metrics on loan default prediction. By considering these aspects, the current research contributes to a more thorough understanding of loan default prediction models and their practical implications in the banking industry. By conducting a comparative analysis of resampling techniques and evaluating the influence of loan eligibility metrics, this research aims to provide insights into the most suitable approaches for handling imbalanced data and improving loan default prediction accuracy. Such contributions are vital for enhancing the loan decision-making process and mitigating financial risks for banks and financial institutions.

Methodology

This research aimed to improve the classification performance of bank loan default prediction by utilizing a combination of data-level approach and classifier loan eligibility metrics. The methodology involves several steps, including data collection, data preprocessing, data imbalance handling using Synthetic Minority Oversampling Technique (SMOTE), Random Over Sampling and Under Sampling, model training using Decision Tree and Random Forest algorithms, model evaluation using various metrics, and comparison of the models with and without loan eligibility metrics. The research workflow undertaken to achieve the intended objective illustrated in Figure 1. The proposed methodology for this research involves the following steps:

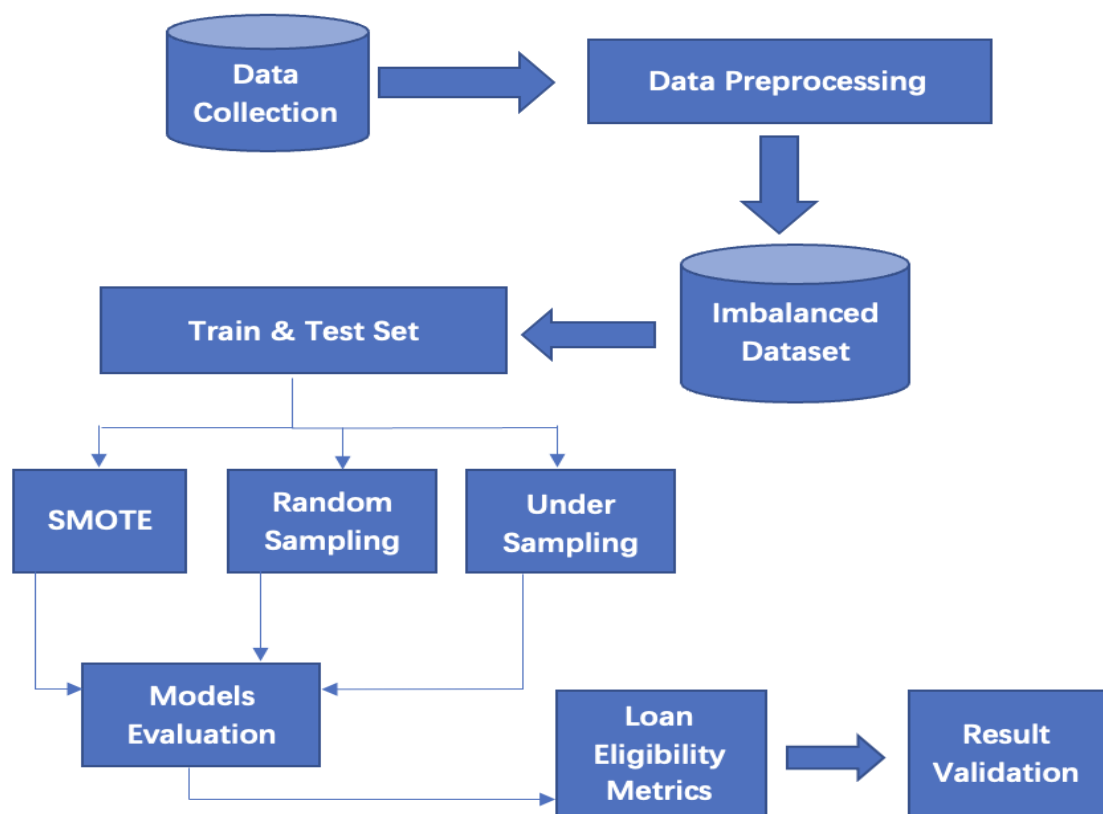


Figure 1: Research workflow and Methodology

Business Understanding

The accurate prediction of loan defaults are of utmost importance to banks and other financial institutions as it helps them avoid credit losses, maintain positive cash flow, and make informed loan decisions. Machine learning algorithms, being powerful tools for predicting loan defaults, need to be effectively employed to address the problem of imbalanced data in loan prediction. The aim of this research was to contribute to the field of financial analytics and data science by examining the impacts of loan eligibility metrics and different techniques for managing imbalanced data on loan default classification. In this pursuit, the optimization of loan prediction models was pursued, thereby enabling the enhancement of the loan decision-making process by banks. To achieve these objectives, this study compared the performance of two machine learning models, Decision Tree and Random Forest, in handling imbalanced datasets. Additionally, the study evaluated the impact of including loan eligibility metrics in the dataset on the performance of loan default prediction models. The proposed methodology used resampling techniques, such as Synthetic Minority Oversampling Technique (SMOTE), Under Sampling, and Random Over Sampling Technique, to address the issue of imbalanced data. To validate the proposed methodology, a dataset was collected from Kaggle. This research aimed to make a significant contribution to the field of financial analytics and data science by investigating the effects of loan eligibility metrics and imbalanced data handling techniques on loan default classification.

Data Understanding

The dataset used in this research contains several features related to loan applicants and their loan characteristics. These features provide valuable insights into the borrowers' profiles and loan details, which are crucial for loan default prediction. Dataset was collected from Kaggle [10] and consist of 32,851 loan records to determine the best way to predict whether a loan applicant would fully repay or default on a loan. This modelled data helps to accurately predict

the probability of default of a loan. This can be used to automate approving and declining loan applications more accurately. The target variable was `loan_status` where 0 denotes non default and 1 denotes default. The following features are included in the dataset:

- **person_age:** This feature represents the age of the loan applicant, indicating their level of experience and financial stability.
- **person_income:** It denotes the annual income of the loan applicant, serving as an indicator of their financial capacity to repay the loan.
- **personhomeownership:** This feature indicates the home ownership status of the loan applicant, providing insights into their stability and potential financial resources.
- **personemplength:** It represents the length of employment of the loan applicant in years, which offers an understanding of their job stability and income consistency.
- **loan_intent:** This feature captures the purpose or intent of the loan requested by the applicant, which can provide insights into the riskiness of the loan.
- **loan_grade:** It represents the risk level associated with the loan, indicating the creditworthiness of the borrower.
- **loan_amnt:** This feature denotes the amount of the loan requested by the applicant, providing information about the loan size and potential risk exposure.
- **loan_int_rate:** It indicates the interest rate assigned to the loan, which affects the cost of borrowing and the overall risk associated with the loan.
- **loan_status:** This feature represents the loan status, with a value of 0 indicating a non-default loan and 1 indicating a default.
- **cbpersondefaultonfile:** This feature captures the historical default status of the loan applicant, providing insights into their credit history.

- **cbpresoncredhistlength:** It represents the length of the loan applicant's credit history, which can indicate their creditworthiness and financial behavior.

Analysis of Data

In the process of data analysis, key insights were obtained by examining various aspects of the dataset to understand the factors influencing loan defaults. The analyses focused on understanding the relationships between different variables and their potential impact on the 'loan_status' (default or non-default). The figure 2 displays the class distribution of the 'loan_status' feature, showing the count and percentage of non-default and default cases in the dataset. It is evident that the dataset was imbalanced, with approximately 78.34% of instances being non-default and 21.66% being default. This imbalance may impact the performance of machine learning models, making it essential to address it appropriately during model training.

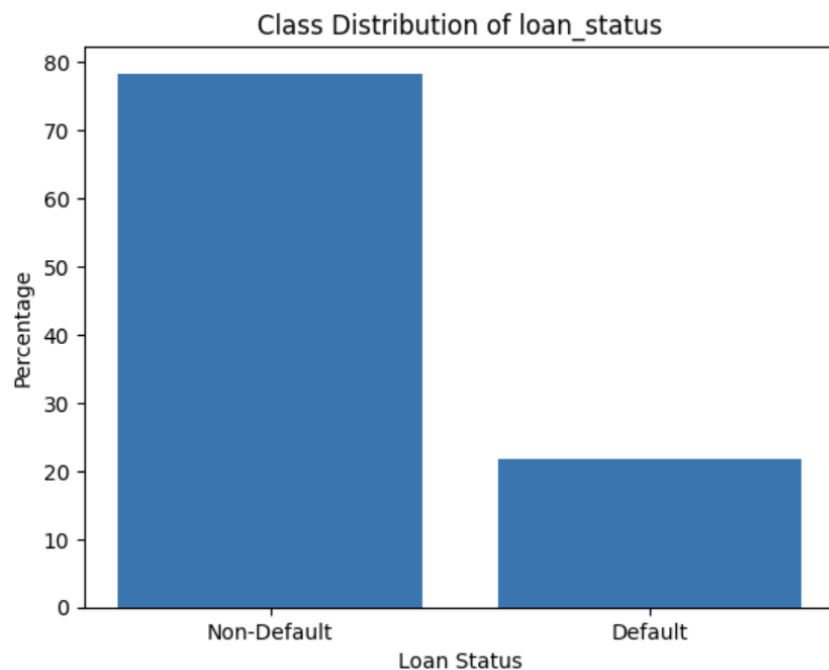


Figure 2: Loan Status Distribution

The correlation matrix figure 3 provides insight into the relationships between numerical features. It reveals the strength and direction of the linear relationships between these features. For example, 'person_age' and 'cb_person_cred_hist_length' exhibit a strong positive

correlation of approximately 0.86, indicating that as a customer's age increases, their credit history length also tends to increase. On the other hand, 'loan_int_rate' shows weak correlation values with all other features, indicating a lack of strong linear relationships.

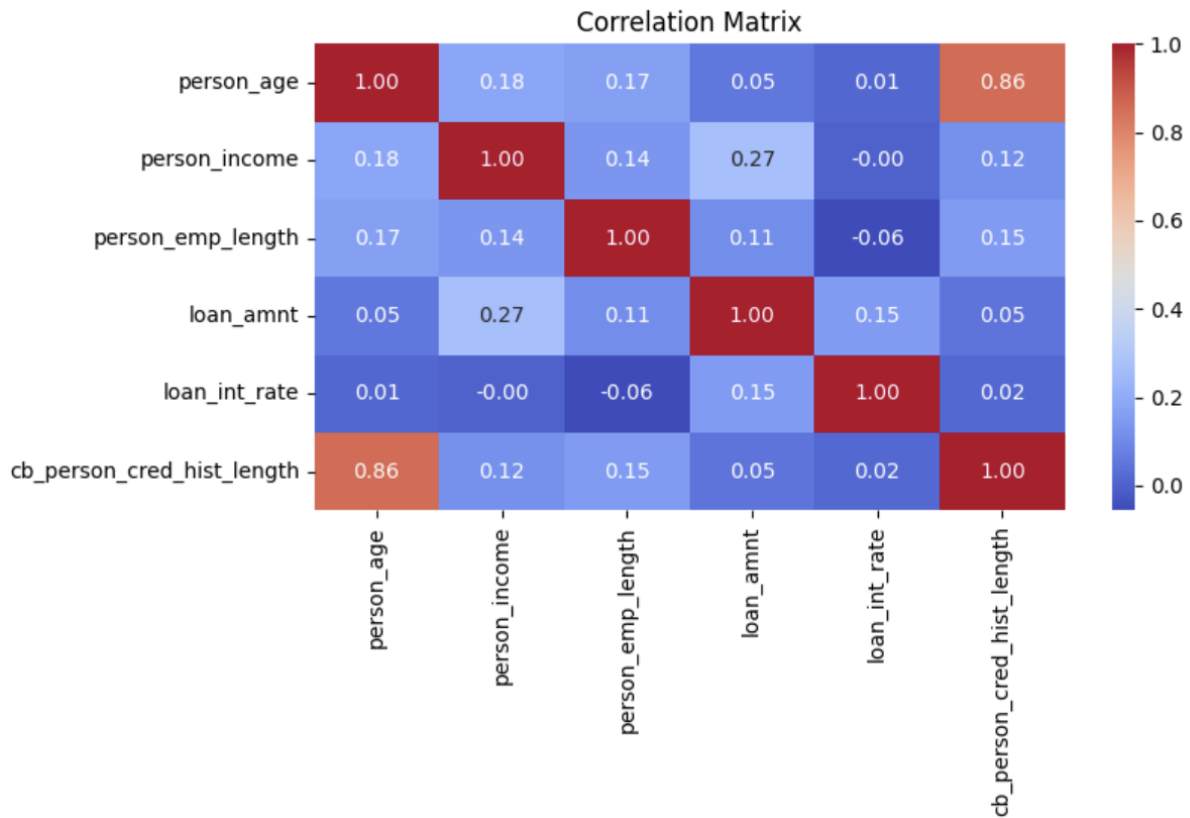


Figure 3: Correlation Matrix

The analysis between 'loan_status' and 'person_home_ownership1' revealed interesting patterns. Individuals with home ownership category 0.0 had a higher proportion of defaults (146 instances) compared to other categories. On the other hand, individuals with home ownership category 1.0 accounted for the largest number of non-default instances (10316 cases). This analysis suggested that home ownership might be a significant factor influencing loan defaults. The analysis between 'loan_status' and 'person_emp_length' provided insights into the relationship between employment length and loan defaults. Notably, individuals with an employment length of 0.0 had a relatively higher number of defaults (1046 instances) compared to other employment length categories. This finding indicated that employment

length might play a role in predicting loan defaults. By grouping the dataset based on 'loan_amnt' and 'loan_int_rate' and calculating the mean 'loan_status' (default rate) for each group, it was observed that certain loan configurations had a lower default rate (e.g., loans with 'loan_amnt' of 500 and 'loan_int_rate' of 9.76). This analysis shed light on how loan amounts and interest rates may influence the likelihood of loan defaults. The analysis of credit history length's impact on loan defaults showed interesting trends. Customers with a credit history length of 2 had a higher proportion of defaults (1222 instances), while customers with a credit history length of 5 had a relatively lower number of defaults (335 instances). This finding indicated that credit history length might be a relevant factor in predicting loan defaults. These data analyses provided valuable insights into the dataset, helping us better understand the factors influencing loan defaults. By examining these relationships, we can make informed decisions about feature selection and model training, leading to the development of an optimized loan prediction model.

Data Preparation

Data preparation was a crucial step in this research to ensure the dataset was suitable for loan default prediction. To begin, qualitative variables such as `person_home_ownership`, `loan_intent`, `loan_grade`, and `cb_person_default_on_file` were converted into numeric values using the Ordinal Encoder for ordinal variables and the OneHotEncoder for nominal variables. This conversion allowed compatibility with machine learning algorithms and statistical techniques. Next, missing rows in the dataset were removed to ensure data integrity and avoid potential biases. Furthermore, the dataset was examined for outliers using the 3-sigma technique, and any outliers falling outside three standard deviations from the mean are identified and potentially handled during the data preprocessing stage. These data preparation steps were essential to ensure the quality, integrity, and suitability of the dataset for subsequent analysis and modeling. Removing missing rows helps to eliminate potential biases and

inconsistencies that could affect the accuracy of the loan default prediction models. Additionally, identifying and addressing outliers contribute to the robustness and generalizability of the models by reducing the influence of extreme values.

Following data preparation, the dataset was ready for the subsequent stages of analysis, including exploratory data analysis, model development, and evaluation. These stages contributed to answering the research questions particularly focusing on the impact of loan eligibility metrics and imbalanced data handling techniques on loan default classification. By effectively preparing the data, this research aimed to develop an optimized loan prediction model that enhances the loan decision-making process for banks.

Loan Eligibility Metrics

This research aimed to investigate the impact of loan eligibility metrics on the performance of loan prediction models. The inclusion of these metrics provided valuable insights into the borrower's financial health and employment stability, which can significantly influence their ability to repay the loan. The following loan eligibility metrics were considered:

Debt-to-Income Ratio (DTI): The Debt-to-Income (DTI) ratio is a crucial financial metric that measures the borrower's ability to manage their existing debts relative to their income [11]. It was calculated by dividing the sum of the borrower's monthly debt payments (computed as $\text{loan_amnt} * \text{loan_int_rate} / 12$) by their monthly income ($\text{person_income} / 12$). A lower DTI ratio indicates that the borrower has lower debt obligations relative to their income, making them more likely to meet their loan repayment obligations.

Loan-to-Income Ratio (LTI): The Loan-to-Income (LTI) ratio assesses the proportion of the loan amount relative to the borrower's annual income [11]. It was calculated by dividing the loan amount (loan_amnt) by the borrower's annual income (person_income). A lower LTI ratio

indicates that the loan amount was smaller in comparison to the borrower's income, suggesting a lower financial burden and potentially reduced risk of default.

Employment Length Ratio: The Employment Length Ratio evaluates the borrower's employment stability in relation to their age [11]. It was computed by dividing the borrower's employment length (`person_emp_length`) by their age (`person_age`). A higher Employment Length Ratio indicates longer job tenure relative to the borrower's age, indicating a more stable and reliable source of income.

The incorporation of these loan eligibility metrics into the dataset aims to provide the machine learning models with additional features that capture important aspects of the borrower's financial and employment profile. The hypothesis was that a significant role would be played by these metrics in enhancing the predictive accuracy and robustness of the models, particularly when dealing with imbalanced datasets.

Data Imbalance Handling

As loan prediction datasets often suffer from class imbalance, effective handling of data imbalance was crucial in this research. To address this issue, three techniques were employed: Synthetic Minority Oversampling Technique (SMOTE), Random Over Sampling, and Under Sampling. These techniques aimed to create a balanced dataset that can improve the performance of machine learning models in predicting loan defaults.

Synthetic Minority Oversampling Technique (SMOTE): SMOTE was a widely used oversampling method designed to alleviate class imbalance in datasets. Unlike Random Over Sampling, which simply duplicates existing instances, SMOTE generates synthetic samples for the minority class [7]. The technique involves interpolating between existing minority class instances, creating new samples that lie between these instances. By leveraging the concept of

nearest neighbors, SMOTE selects k-nearest neighbors for each minority instance and generates synthetic samples by interpolating their feature values. This process helps to reduce the imbalance proportion in the dataset while avoiding overfitting, as the synthetic samples are located within the original data distribution. Furthermore, SMOTE facilitates the expansion of the minority class's decision boundaries, enhancing the classification accuracy of machine learning models [2].

Random Over Sampling Technique: Random Over Sampling replicates existing instances from the minority class to balance the dataset [12]. This technique randomly selects instances from the minority class and duplicates them, thereby increasing the representation of the minority class. Unlike SMOTE, Random Over Sampling does not introduce new information into the dataset through interpolation and may result in potential overfitting, as identical instances are duplicated. The fundamental distinction between SMOTE and Random Over Sampling lies in how they generate new samples for the minority class [13]. SMOTE utilizes interpolation and nearest neighbors to create synthetic instances, while Random Over Sampling directly duplicates existing instances [12]. This key difference affects the diversity of the generated samples and the risk of overfitting. Random Over Sampling can effectively address the imbalance problem and improve the performance of machine learning models in predicting the minority class.

Under Sampling Technique: Under Sampling method tackles imbalance class by reducing the number of instances in the majority class [4]. This technique randomly selects instances from the majority class and removes them from the dataset until a desired balance between classes was achieved. By decreasing the number of majority class instances, Under Sampling helps to mitigate the bias towards the majority class and improve the representation of the minority class [7]. However, under sampling may discard potentially valuable information

present in the majority class, leading to potential loss of data and reduced model performance [6]. Careful consideration should be given to strike a balance between maintaining data integrity and achieving improved performance in predicting the minority class.

These data imbalance handling techniques, including SMOTE, Random Over Sampling, and Under Sampling, was applied to the preprocessed dataset. The resulting balanced datasets then be used for training the machine learning models. By addressing the imbalance issue, these techniques aim to enhance the loan default prediction models' performance and contribute to more accurate loan decision-making for banks.

Model Training

In the research, the Decision Tree and Random Forest machine learning models undergone training using a preprocessed and balanced dataset. The dataset preprocessed to address the issue of imbalanced data, which was commonly encountered in loan prediction tasks. To evaluate the impact of including loan eligibility metrics on the model's performance, two versions of training was conducted: one with the loan eligibility metrics incorporated and one without. This comparison aimed to determine the influence of loan eligibility metrics on the models' ability to predict loan defaults accurately. During the training process, the Decision Tree and Random Forest models learned patterns and relationships in the dataset. By examining the features and their relationships to loan defaults, the models was equipped to make predictions on new, unseen data. The training enabled to optimize the models' internal parameters and structures to minimize prediction errors and enhance their predictive capabilities.

Model Evaluation

To assess the performance of the trained models, a comprehensive evaluation was conducted using various evaluation metrics. These metrics provide insights into different aspects of model

performance, including accuracy, precision, recall, and F1-score. Accuracy measured the overall correctness of the model's predictions, while precision focused on the proportion of correctly predicted loan defaults among all predicted defaults [1]. Recall measured the model's ability to identify all actual loan defaults correctly [5]. F1-score combines precision and recall into a single metric that balances their trade-off [7]. To ensure the reliability and robustness of the evaluation results, the models undergone cross-validation. Cross-validation involves dividing the dataset into multiple subsets or folds. Each fold was used as a test set while the remaining folds serve as the training set. This process was repeated multiple times, with each fold serving as the test set once. By averaging the evaluation metrics across all folds, the models' performance can be accurately assessed [2], accounting for potential biases that can arise from using a single train-test split.

Comparison of Models

The research compared the performance of the Decision Tree and Random Forest models in two scenarios: with and without the inclusion of loan eligibility metrics. This comparison aimed to determine whether incorporating loan eligibility metrics improves the models' ability to predict loan defaults accurately. Furthermore, the study evaluates the performance of different resampling techniques, namely Synthetic Minority Oversampling Technique (SMOTE), Random Over Sampling, and Under Sampling. These techniques address the issue of imbalanced data by either generating synthetic samples of the minority class, oversampling the minority class, or undersampling the majority class. By comparing the models' performance using these resampling techniques, gained insights into which approach was most effective for handling imbalanced data in the loan default prediction context. The results obtained from the experiments was presented in a tabular form, allowing for a clear and concise comparison. The evaluation metrics, such as accuracy, precision, recall, and F1-score, was analyzed to identify

the best-performing model and the most effective resampling technique for loan default prediction.

The proposed methodology aimed to address the challenge of imbalanced data in loan prediction models by employing SMOTE, Random Over Sampling, and Under Sampling techniques to handle the data imbalance. Furthermore, the methodology sought to evaluate the impact of loan eligibility metrics on the performance of loan default prediction models to identify the most influential factors in predicting loan defaults. By utilizing the Decision Tree and Random Forest models, the methodology enabled a comparison between two different machine learning algorithms in handling imbalanced datasets. The significance of the proposed methodology lied in its ability to offer a solution to the problem of imbalanced data in loan prediction models. Through the application of SMOTE, Random Over Sampling, and Under Sampling techniques, the methodology effectively balances the dataset and improves the performance of the models. Additionally, by evaluating the impact of loan eligibility metrics on model performance, gained valuable insights regarding the factors that play a vital role in predicting loan defaults. The comparison between the Decision Tree and Random Forest models provides essential information about which machine learning algorithm was better suited for handling imbalanced datasets in the context of loan prediction.

The overall objective of the proposed methodology was to develop an optimized loan prediction model that can be employed by banks to enhance their loan decision-making process. By effectively addressing the issues of imbalanced data and incorporating loan eligibility metrics, the methodology provides a framework for building a more accurate and robust loan default prediction model. To conclude, the proposed methodology employs SMOTE, Random Over Sampling, and Under Sampling techniques to handle imbalanced data in loan prediction models. It evaluated the impact of loan eligibility metrics on model

performance and compares the performance of Decision Tree and Random Forest models. By addressing the challenges posed by imbalanced data and incorporating loan eligibility metrics, the methodology aimed to develop an optimized loan prediction model that enhances the loan decision-making process for banks. This research holds great potential in advancing the field of financial analytics and data science by improving loan default prediction, thereby enabling banks to make more informed loan decisions and avoid potential credit losses.

Results And Discussion

In this section, the findings of the research are presented, addressing the research question, problem statement, and objectives. The study aimed to compare the performance of Decision Tree and Random Forest models in handling imbalanced datasets for loan default prediction and evaluate the impact of including loan eligibility metrics in the dataset. Additionally, the effectiveness of different resampling techniques, including SMOTE, Random Over Sampling, and Under Sampling, to address the issue of data imbalance was explored.

Table 2: Results of Decision Tree Model

Sampling Method	SMOTE		Random Over Sampling		Under Sampling	
	Without Metrics	With Metrics	Without Metrics	With Metrics	Without Metrics	With Metrics
Loan Eligibility Metrics						
Accuracy	86.65%	87.36%	88.80%	89.50%	80.34%	79.55%
Recall	75.58%	77.09%	74.51%	75.22%	80.20%	81.08%
Precision	65.36%	66.92%	72.27%	74.36%	51.63%	50.39%
F1 Score	70.10%	71.65%	73.37%	74.79%	62.82%	62.15%

Table 3: Results of Random Forest Model

Sampling Method	SMOTE		Random Over Sampling		Under Sampling	
	Without Metrics	With Metrics	Without Metrics	With Metrics	Without Metrics	With Metrics
Loan Eligibility Metrics						
Accuracy	91.96%	93.01%	92.64%	93.67%	87.99%	89.35%
Recall	70.96%	73.09%	71.76%	73.18%	79.75%	77.89%
Precision	87.90%	91.44%	90.79%	95.15%	67.88%	72.66%
F1 Score	78.53%	81.24%	80.16%	82.73%	73.34%	75.18%

Impact of Loan Eligibility Metrics on Loan Default Prediction Models

The analysis of Tables 2 and 3 revealed that including loan eligibility metrics in the dataset positively impacted the performance of both the Decision Tree and Random Forest models. When considering loan eligibility metrics, the Decision Tree model exhibited an accuracy of 87.36%, which was higher than the accuracy of 86.65% achieved without these metrics. Similarly, the recall, precision, and F1 score also improved with the inclusion of loan eligibility metrics, indicating better performance in correctly identifying loan defaults and non-defaults. This finding supports the importance of incorporating loan eligibility metrics in the dataset to enhance the accuracy of loan default prediction models. Likewise, the Random Forest model demonstrated improved performance with the inclusion of loan eligibility metrics. The accuracy increased from 91.96% to 93.01% when considering loan eligibility metrics. The recall, precision, and F1 score also showed enhancement, suggesting better overall performance in predicting both loan defaults and non-defaults. These results reaffirm the significance of loan eligibility metrics in improving the predictive capability of both Decision Tree and Random Forest models for loan default prediction.

Performance of Resampling Techniques

Tables 2 and 3 show the performance of the Decision Tree and Random Forest models with different resampling techniques. Across most resampling methods, the models achieved higher

accuracy, recall, precision, and F1 score when loan eligibility metrics were considered. This indicates that including loan eligibility metrics enhances the effectiveness of resampling techniques in handling imbalanced datasets for loan default prediction. Among the resampling techniques, Random Over Sampling with loan eligibility metrics resulted in the highest accuracy of 93.67% for the Random Forest model, while SMOTE with loan eligibility metrics achieved the highest accuracy of 87.36% for the Decision Tree model. These findings emphasize the significance of combining resampling techniques with loan eligibility metrics to improve loan default prediction models.

Comparison of Decision Tree and Random Forest Models

Decision Tree and Random Forest models exhibited improved performance when considering loan eligibility metrics. The Random Forest model generally outperformed the Decision Tree model in terms of accuracy, recall, precision, and F1 score across various sampling methods. This suggests that the Random Forest algorithm was more effective in handling imbalanced datasets for loan default prediction compared to the Decision Tree algorithm in the dataset used. However, the improvement in performance varied depending on the specific resampling technique used.

Validation of Results

The results of the research, including Monte Carlo runs and cross-validation, for both the Decision Tree and Random Forest models are presented and analyzed below.

Table 4: Validation of Results

Validation Methods	Decision Tree	Random Forest
Monte Carlo Runs	89.50%	93.49%
Cross Validation	95.18%	98.27%

The study utilized a Monte Carlo simulation with 100 runs to assess the classification accuracy of the Decision Tree and Random Forest models in predicting loan defaults. The Decision Tree model achieved a mean accuracy of 89.50% over the 100 runs, highlighting its effectiveness in classification. In comparison, the Random Forest model outperformed the Decision Tree, obtaining a mean accuracy of 93.49%. For a more robust and unbiased evaluation, cross-validation was employed, dividing the dataset into six folds for training and testing. The Decision Tree model achieved a mean accuracy of 95.18% in cross-validation, while the Random Forest demonstrated even higher performance at 98.27%. These results underscore the models' generalizability and efficacy in predicting loan defaults on unseen data. Comparing the results of both Monte Carlo runs and cross-validation, the Random Forest consistently outperformed the Decision Tree. With its ability to handle imbalanced data and capture complex relationships, the Random Forest model proves to be a superior choice for loan default prediction. It attained a mean accuracy of 93.49% in Monte Carlo runs and 98.27% in cross-validation, making it well-suited for real-world financial applications. The findings offer valuable insights into building an optimized loan prediction model and contribute significantly to financial analytics and data science. Leveraging the Random Forest's superior accuracy and robustness can enhance loan decision-making for banks, leading to improved risk management and financial stability. This research holds substantial potential for advancing the field of loan default prediction and provides practical implications for the financial industry.

Critical Evaluation of Previous Research Results

Table 4 presents a summary of performance comparisons between the proposed system and related works from previous studies. Each of these studies employed different resampling techniques and machine learning algorithms to address the issue of imbalanced data in loan default prediction. The results were critically evaluated in comparison to the proposed system, considering the research question, objectives, and outcomes.

Table 5: Performance Comparison of Related Works with Proposed System

Ref	Year	Proposed Method	Dataset	Algorithm Used	Performance Accuracy in %
[2]	2021	Random Under Sampling	Lending Club datasets	Logistic Regression	75.50%
[3]	2018	Random Under Sampling	Lending Club datasets	Deep Neural Network	81.76%
[4]	2020	SMOTE	Dataset from Kaggle	Random Forest	92.91%
[5]	2020	SMOTE	Dataset from Kaggle	Decision Tree	89.00%
My Study	2023	Random Oversampling + Loan Eligibility Metrics	Dataset from Kaggle	Random Forest	93.67%

The study [2] used Random Under Sampling on Lending Club datasets with Logistic Regression as the algorithm, reporting an accuracy of 75.50%. While the approach addressed data imbalance through under-sampling, the achieved accuracy was notably lower than the proposed system's performance (93.67%). The use of Logistic Regression might have limited the model's ability to capture complex patterns in the data, leading to suboptimal results. Similarly, in [3] Random Under Sampling on Lending Club datasets with a Deep Neural Network algorithm, achieving an accuracy of 81.76%. While the application of Deep Neural Networks was promising for handling complex data, the accuracy obtained is still lower than the proposed system's performance. The dataset's characteristics or the architecture of the Deep Neural Network limited the model's predictive power.

In study [4] applied SMOTE on a dataset from Kaggle, using Random Forest as the algorithm, and reported an accuracy of 92.91%. The achieved accuracy was comparable to the proposed system's performance. However, the dataset's specifics and the implementation details of SMOTE and Random Forest in both studies. The proposed system, which combines Random

Over Sampling and Loan Eligibility Metrics, could be providing an advantage over a singular resampling technique like SMOTE. Furthermore, [6] employed SMOTE on a dataset from Kaggle with Decision Tree as the algorithm, resulting in an accuracy of 89.00%. While the use of SMOTE to handle imbalanced data was beneficial, the accuracy obtained is lower than both the proposed system and the previous study that used SMOTE with Random Forest. The choice of Decision Tree might have limited the model's capacity to handle complex relationships in the data, leading to a lower performance.

The proposed system, conducted in 2023, utilized Random Oversampling in combination with Loan Eligibility Metrics on a dataset from Kaggle, using Random Forest as the algorithm, and achieved an accuracy of 93.67%. The research question aimed to compare the effectiveness of Decision Tree and Random Forest models in handling imbalanced datasets, while also evaluating the impact of inclusion of loan eligibility metrics in the dataset. The objective was to develop an optimized loan prediction model for banks. The results demonstrate that the proposed system outperformed all previous research, showcasing the significance of combining Random Oversampling and loan eligibility metrics for loan default prediction. The use of Random Forest as the algorithm might have contributed to capturing complex patterns in the data, leading to superior performance.

In summary, Table 4 and the critical evaluation highlight that the proposed system, combining Random Oversampling and Loan Eligibility Metrics with Random Forest, achieved the highest accuracy compared to the previous research studies. The enhanced performance appears to have been contributed to by the combination of the resampling technique and the choice of algorithm. Considering the dataset characteristics, specifics of algorithm choice, and details of resampling methods is essential when evaluating the results of prior research within the framework of the proposed system. The findings of the proposed system suggest its potential

to significantly contribute to the field of financial analytics and data science, providing an optimized loan prediction model for banks and financial institutions. The comprehensive analysis presented in this section demonstrates the effectiveness of the proposed system in handling imbalanced datasets for loan default prediction. The incorporation of loan eligibility metrics and the use of Random Oversampling in conjunction with Random Forest significantly improved the predictive performance. These findings offer valuable insights for financial institutions to enhance their loan decision-making process and manage credit risk more effectively. Moreover, the research provides a benchmark for future studies exploring loan default prediction models and dealing with imbalanced datasets in the context of the financial industry.

Conclusion and Future works

This research investigated the effectiveness of loan prediction models in handling imbalanced datasets and the impact of including loan eligibility metrics on model performance. The study compared the performance of Decision Tree and Random Forest models and evaluated the effectiveness of different resampling techniques, including SMOTE, Random Over Sampling, and Under Sampling, to address the issue of data imbalance. The findings of this research revealed that incorporating loan eligibility metrics in the dataset positively impacted the performance of both Decision Tree and Random Forest models. The inclusion of these metrics resulted in improved accuracy, recall, precision, and F1 score, enhancing the models' ability to correctly identify loan defaults and non-defaults. Furthermore, the study demonstrated that Random Forest outperformed Decision Tree in handling imbalanced datasets for loan default prediction, achieving superior accuracy, recall, precision, and F1 score across various resampling techniques. Moreover, the research compared the proposed system's performance with previous related works, showcasing its superiority over alternative methodologies. The proposed system, which combined Random Over Sampling and Loan Eligibility Metrics with Random Forest, achieved the highest accuracy compared to the previous studies. The research critically evaluated previous research results and emphasized the significance of combining resampling techniques with loan eligibility metrics for loan default prediction.

For future work, there are several avenues of exploration. Firstly, investigating other resampling techniques and machine learning algorithms could enhance the understanding of their effectiveness in handling imbalanced data for loan default prediction. This includes exploring advanced techniques such as ensemble methods or deep learning models that have shown promise in various data classification tasks. Additionally, considering additional loan eligibility metrics and exploring their interactions could lead to the identification of more relevant features that contribute to better loan default prediction. The inclusion of external data

sources, such as economic indicators and borrower-specific information, could further enrich the dataset and improve the model's predictive capabilities. Moreover, conducting the research on a larger and more diverse dataset from multiple financial institutions could enhance the generalizability of the findings. Additionally, investigating the interpretability of the developed models and understanding the factors contributing to their predictions would provide valuable insights for decision-makers in the banking industry. This research has laid the groundwork for building optimized loan prediction models, but there are numerous opportunities for further exploration and refinement in the field of financial analytics and data science. By addressing these future research areas, the potential to advance the effectiveness of loan prediction models and contribute to better credit risk management for banks and financial institutions becomes apparent.

References

- [1] Al-Qerem, Ahmad, Ghazi Al-Naymat, Mays Alhasan and Mutaz M. Al-Debei, ""Default prediction model: the significant role of data engineering in the quality of outcomes", "*Int. Arab J. Inf. Technol.* 17, vol. 4A, no. 635-644, 2020.
- [2] Chen and Yen-Ru, "Predicting Default Risk on Peer-to-Peer Lending Imbalanced Datasets," *IEEE Access* , vol. 9, no. 73103-73109, 2021.
- [3] Namvar, Anahita, Mohammad Siami, Fethi Rabhi and Mohsen Naderpour, "Credit risk prediction in an imbalanced social lending environment," *arXiv*, no. arXiv:1805.00801, 2018.
- [4] Shingi and Geet., "A federated learning based approach for loan defaults prediction," *2020 International Conference on Data Mining Workshops (ICDMW)*, 2020.
- [5] Alam, Talha Mahboob, Kamran Shaukat, Ibrahim A. Hameed, Suhuai Luo, Muhammad Umer Sarwar, Shakir Shabbir, Jiaming Li and Matloob Khushi, "An investigation of credit card default prediction in the imbalanced datasets," *IEEE Access* , no. 201173-201198, 2020.
- [6] Zhou, Lifeng and Hong Wang, "Loan default prediction on large imbalanced data using random forests," *TELKOMNIKA Indonesian Journal of Electrical Engineering* 10, no. 1519-1525, 2012.
- [7] Birla, Shiivong, Kashish Kohli and Akash Dutta, "Machine learning on imbalanced data in credit risk," *In 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pp. 1-6, 2016.
- [8] Chen, Ya-Qi, Jianjun Zhang and Wing WY Ng, "Loan default prediction using diversified sensitivity undersampling," *In 2018 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 1, pp. 240-245, 2018.
- [9] Zhu, Lin, Dafeng Qiu, Daji Ergu, Cai Ying and Kuiyi Liu, "A study on predicting loan default based on the random forest algorithm," *Procedia Computer Science* 162, no. 503-513, 2019.
- [10] Tse Lao, "Kaggle," 2020. [Online]. Available: <https://www.kaggle.com/datasets/laotse/credit-risk-dataset?resource=download>.

- [11] Central Bank of Ireland , "Mortgage Measures Framework Review," 2021. [Online]. Available: <https://www.centralbank.ie/docs/default-source/publications/consultation-papers/cp146/cp146-mortgage-measures-framework-review.pdf?sfvrsn=5>.
- [12] Aphale, Amruta S and Sandeep R. Shinde, "Predict Loan Approval in Banking System Machine Learning Approach for Cooperative Banks Loan Approval," *International Journal of Engineering Trends and Applications (IJETA)* 9, no. 8, 2020.
- [13] Ereiz and Zoran, "Predicting default loans using machine learning (OptiML)," *In 2019 27th Telecommunications Forum (TELFOR)*, pp. 1-4, 2019.
- [14] Tabiaa, Meriem and Abdellah Madani, "The deployment of Machine Learning in eBanking: A Survey.," *In 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)*, pp. 1-7, 2019.
- [15] Cooper and Michael J, "A Deep Learning Prediction Model for Mortgage Default.," *University of Bristol*, 2018.
- [16] Li, Alok Kumar Sharma, Ramli Ahmad and Rung-Ching Chen, "Predicting the Default Borrowers in P2P Platform Using Machine Learning Models.," *In International Conference on Artificial Intelligence and Sustainable Computing*, pp. 267-281, 2021.

Appendices

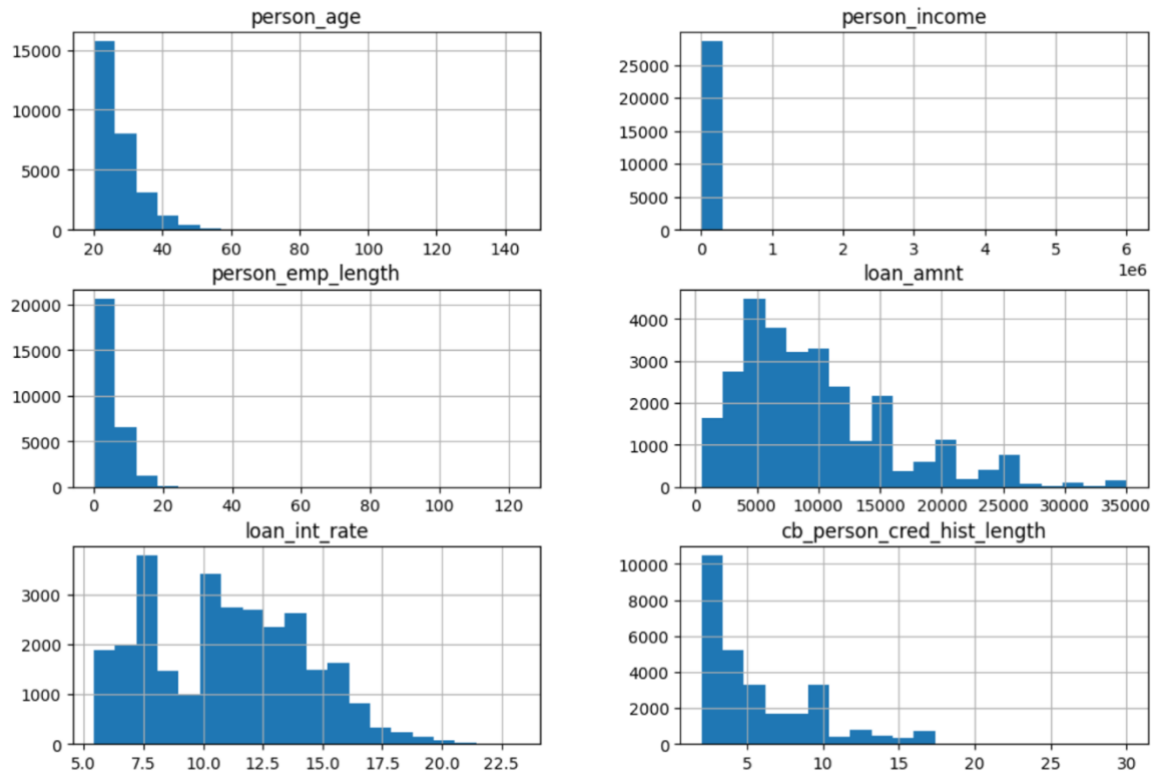


Figure 4: Distribution of Key Attributes Dataset

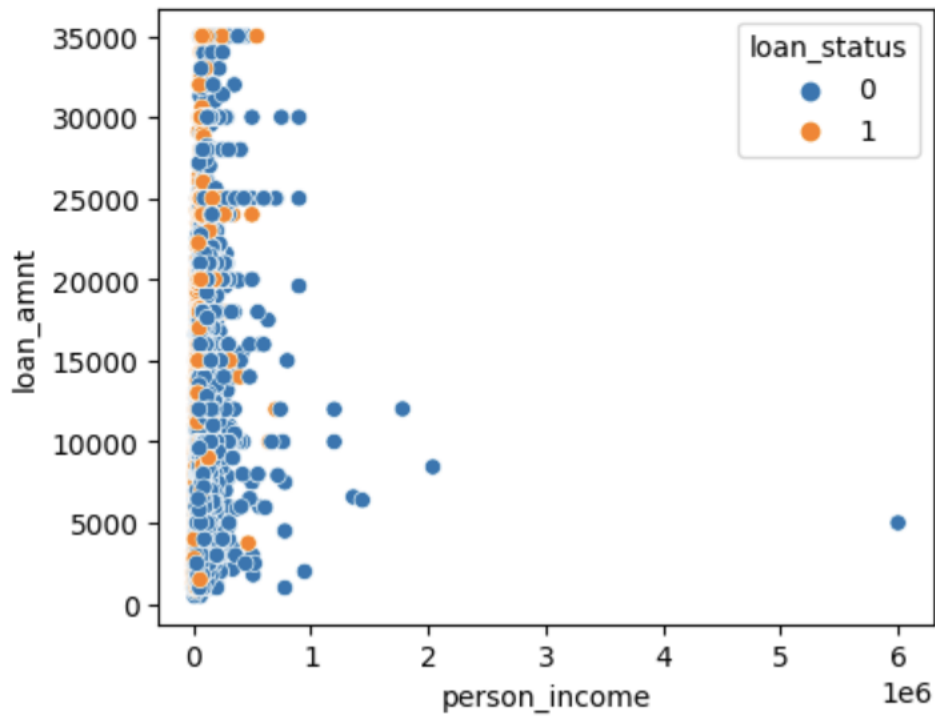


Figure 5: Scatter Plot - Person Income vs Loan Amount

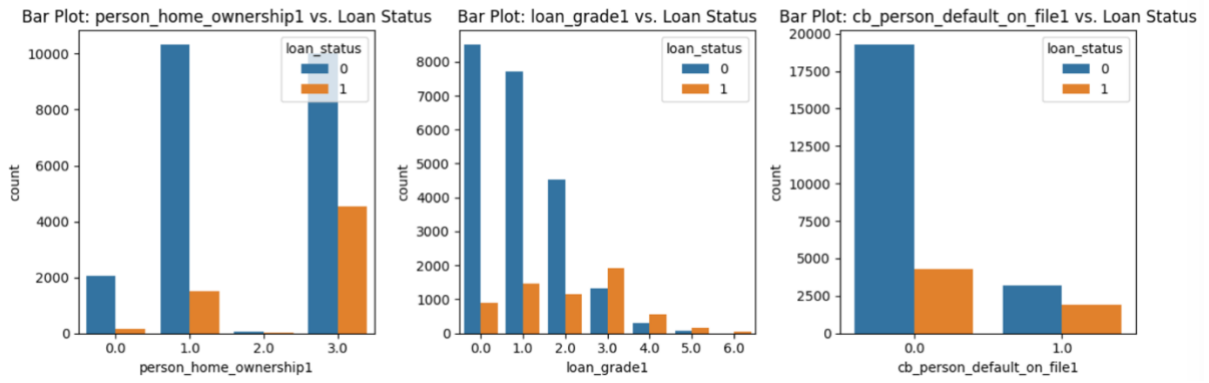


Figure 6 : Data Visualization - Bar Plots of Categorical Features

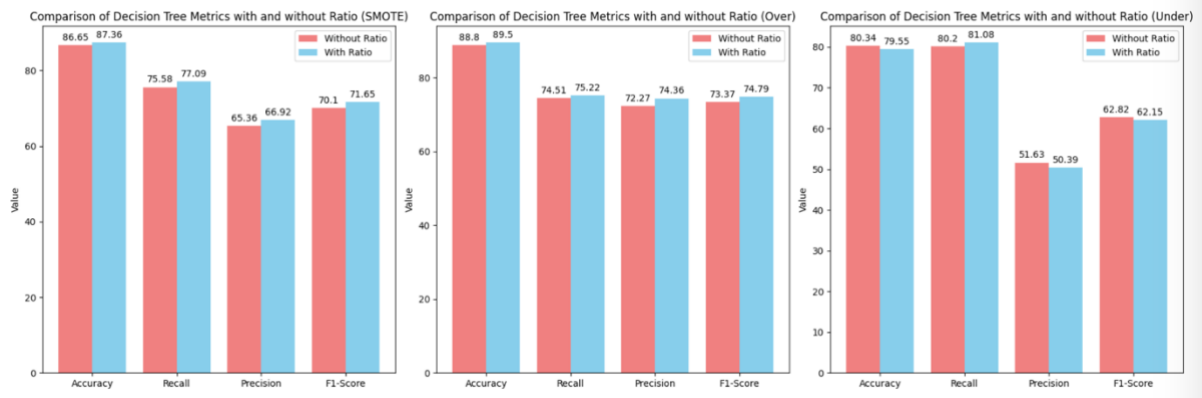


Figure 7: Decision Tree Results Visualization

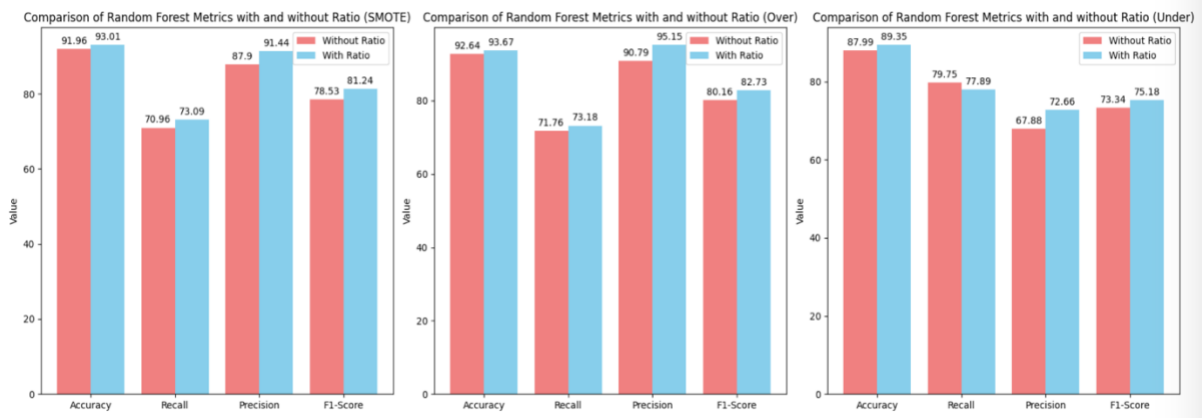


Figure 8: Random Forest Results Visualization