Comparing Loan Default Predictions Using Machine Learning:

A Comparative Approach to Model Selection and Data Balancing Methods

Joseph Do

STUDENT NUMBER: 24039873

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN BUSINESS ANALYTICS

KENT BUSINESS SCHOOL

UNIVERSITY OF KENT

Supervisor: Dr Mingzhe Wei

Word count: 8,790

University of Kent

Kent Business School

Kent, The United Kingdom

August 2024

Canterbury, Kent, United Kingdom, 2024

Acknowledgements

This project was a meaningful part of my studies at the University of Kent. It brought together everything I learned in my MSc Business Analytics course. It took a lot of time and was quite challenging, but it was a unique experience that I will always remember fondly. I want to thank the people who helped me complete it successfully:

- Firstly, I want to thank my family, including my parents and brother, who provided everything—money, time, and support—so I could get a good education in this country. They have always been by my side, encouraging me to overcome challenges.
- Next to, I want to thank all the lecturers who shared their valuable knowledge with me. I am especially grateful to my supervisor, Dr Mingzhe Hei, for his guidance and support.
- I want to thank all my friends who helped and encouraged me to successfully complete my project. I am very grateful to each of them.

Table of Contents

List of Figures

List of Equations

List of Tables

Abstract

This thesis investigates the effectiveness of machine learning models in predicting non-performing loans (NPLs) in private sector, a critical challenge for financial institutions due to the financial risks associated with loan defaults. The study evaluates three machine learning models—Logistic Regression, Decision Tree, and Random Forest—with a focus on handling imbalanced datasets, a common issue in loan default prediction. To address this imbalance, techniques such as cost-sensitive learning, undersampling, and oversampling were applied and compared.

The results demonstrate that the Random Forest model, particularly when used with oversampling, achieves superior predictive performance across multiple metrics, including accuracy, precision, recall, and F1-Score. This highlights its robustness in distinguishing between defaulted and non-defaulted loans. While cost-sensitive learning provided some improvement, it did not perform as effectively as expected, suggesting the need for more precise cost calibration. The findings of this study offer valuable insights for enhancing credit risk assessment processes and point to the potential of advanced ensemble methods in improving the prediction of loan defaults.

Comparing Loan Default Predictions Using Machine Learning:

A Comparative Approach to Model Selection and Data Balancing

Joseph Do

1. Introduction

Based on the most recent data from Fitch Ratings, the loan default rates in the financial sector have seen a significant increase in 2023, highlighting the growing risks within the market. Specifically, the leveraged loan default rate in the U.S. rose from a relatively low 0.85% in January 2023 to approximately 2.5% by December 2023. This increase is primarily attributed to the challenging macroeconomic conditions, including higher interest rates and tighter credit conditions, which have made it more difficult for borrowers, especially those with speculative-grade credit ratings, to refinance their debts.

Moreover, Fitch Ratings reported that the trailing 12-month default rate for leveraged loans had increased to 3.04% by the end of 2023, up from 1.6% the previous year. They forecast further increases in 2024, predicting that the default rate could rise to between 3.5% and 4% due to continued economic pressures.

These statistics underscore the critical need for more robust and reliable analytical methods in financial risk assessment. The increasing default rates reveal the limitations of current predictive models, which have struggled to adapt to the rapidly changing economic landscape. Consequently, the motivation for this study is to address these inadequacies by exploring and developing more effective tools for forecasting and mitigating credit risk within the financial sector.

In recent years, machine learning has emerged as a powerful tool for improving the accuracy of loan default predictions. Unlike traditional models like Logistic Regression often fall short in capturing the intricate patterns that indicate a borrower's likelihood of default in a more complex dataset, machine learning algorithms can identify complex relationships within large datasets, making them particularly effective for this task. Among the various machine learning techniques, models like Decision Trees and Random Forests have shown promise in handling the non-linear patterns that are often present in financial data.

This thesis explores the application of three machine learning models—Logistic Regression, Decision Tree, and Random Forest—in predicting non-performing loans. The study pays particular attention to the challenge of imbalanced datasets, where defaulted loans are much less common than fully repaid ones. To address this, different strategies such as cost-sensitive learning, undersampling, and oversampling are applied and evaluated.

The structure of this thesis is as follows: Section 2 reviews relevant literature on loan default prediction and the use of machine learning in this context. Section 3 outlines the methodology used for data collection and model evaluation. Section 4 presents the results of the

experiments, including a comparison of model performances. Finally, Section 5 concludes with a summary of findings and suggestions for future research.

2. Literature Reviews

2.1 Non-performing loans

Non-performing loans (NPLs) within the private customer segment are a critical issue that can disrupt both financial institutions and the broader economy. NPLs occur when private borrowers, such as individual consumers, fail to make scheduled interest payments or repay the principal amount on their loans. This situation is particularly concerning because private customers typically represent a large portion of a bank's loan portfolio, making the management of these loans crucial for maintaining financial stability.

High levels of NPLs among private customers can severely constrain a bank's ability to lend. When a significant portion of a bank's assets are tied up in non-performing loans, the bank must set aside more capital to cover potential losses, which reduces the funds available for new lending. This reduction in available credit affects not just businesses but also consumers, who may find it more challenging to obtain loans for major purchases such as homes, vehicles, or education. Consequently, the overall economic activity can slow down, as consumer spending and investment diminish.

Accurately predicting defaults among private customers is essential for both lenders and borrowers. For borrowers, accurate predictions can prevent the financial distress associated with taking on unmanageable debt, reducing the likelihood of personal bankruptcy. For lenders, improved default prediction models allow for better risk management, enabling them to offer loans to a broader range of customers while maintaining financial security. This also helps in keeping interest rates lower, as the risk of defaults decreases.

Empirical evidence highlights the significant impact of NPLs in the private sector. For instance, the International Monetary Fund (IMF) reports that a one percentage point increase in the NPL ratio can lead to a GDP growth reduction of 0.1 to 0.5 percentage points in the following year (IMF, 2016). In Europe, the European Central Bank (ECB) has found that countries with private sector NPL ratios exceeding 5% experienced an average growth reduction of 2% compared to those with lower NPL levels (ECB, 2018). Similarly, in Asian financial markets, economies with persistently high private sector NPL ratios (over 10%) have seen GDP growth decline by an average of 1.5 percentage points over a three-year period (Lee and Yeh, 2019).

The strong correlation between NPLs in private customers and economic performance underscores the importance of robust risk management practices in banking. Implementing effective regulatory frameworks and advanced decision-making tools, such as machine learning and analytics, can significantly mitigate the risks associated with private customer loan defaults. These measures help lenders better assess creditworthiness, thereby reducing the likelihood of loan defaults and ensuring more stable economic conditions. For smaller lending firms, particularly those dealing with customers lacking substantial credit histories, these tools are invaluable in making informed loan approval decisions, ultimately minimizing financial losses and promoting responsible lending practices

2.2 Machine learning with NPLs

Machine learning (ML) models are increasingly being used in various fields, including credit risk assessment. By applying supervised ML techniques, traditional risk assessments could be enhanced as these models provide a more comprehensive analysis of a loan applicant beyond the simple linear evaluations of a few risk factors. Traditional risk scoring models often deliver inconsistent and unreliable results (Ereiz 2019). In order to improve the predictive, commonly used ML algorithms in this domain include logistic regression, decision trees, random forests, support vector machines (SVM), and neural networks (Khandani, Kim & Lo, 2010; Qi, Zhang & Zhao, 2020). This paper explored three ML algorithms for predicting defaults: logistic regression, decision tree, and random forest. However, the highly imbalanced nature of loan data continues to be an issue, with the cost of incorrectly predicting a fully paid loan generally outweighing the benefits of correct predictions.

While logistic regression remains popular for its simplicity and effectiveness in binary classification problems, such as determining the likelihood of a loan default. Decision trees and random forests offer the advantage of handling large datasets with multiple features, making them particularly useful in scenarios with complex interactions among variables (Zhang, 2019). Moreover, neural networks, especially deep learning models, have gained attention for their ability to model intricate, non-linear relationships in data, leading to improved accuracy in predicting defaults (Goodfellow, Bengio & Courville, 2016).

However, despite their potential, ML models face challenges, particularly when dealing with highly imbalanced datasets where non-default cases vastly outnumber default cases. This imbalance can lead to models that are biased towards predicting non-defaults, thus reducing the overall effectiveness of the predictions (Chawla et al., 2002). Additionally, the cost of incorrectly predicting a fully paid loan as a default can outweigh the benefits of correctly predicting actual defaults, necessitating careful calibration of these models (Zhou et al., 2022).

2.3 Objectives

The main objective of this study is to evaluate and compare the effectiveness of different machine learning models in predicting non-performing loans (NPLs) in banking, especially focus on private customers. Specifically, the study focuses on three models: Logistic Regression, Decision Tree, and Random Forest. By comparing these models, the study aims to identify which approach offers the best accuracy and reliability in predicting loan defaults.

Another key objective is to explore how different methods for handling imbalanced datasets, such as cost-sensitive learning, undersampling, and oversampling, affect the performance of these models. Since loan default prediction often involves datasets where default cases are much fewer than non-default cases, addressing this imbalance is crucial for improving prediction accuracy.

The study is structured around the following three research questions, which encapsulate all its objectives:

Research question 1:

'Which machine learning algorithm from the selected group shows the best results in predicting loan defaults, according to specific model evaluation criteria?'

A unique aspect of this thesis is the application of three different models using various methods like cost-insensitive, cost-sensitive, under sampling, and over sampling to explore if the prediction of non-performing loans (NPLs) can be enhanced. This methodological approach to loan data is novel in the literature and significant because it could lower the costs associated with false negatives, given that the potential risk of default can reach up to 100%.

Research question 2:

'How much the ensemble method can improve the accuracy of predicting loan defaults?'

This study aims to explore the extent of improvement that ensemble methods can offer compared to traditional models in predicting loan defaults. Given the intricate nature of these ensemble models, it is crucial to determine whether their potential benefits justify the additional complexity and effort required over simpler, conventional models.

Research question 3:

"Which method for managing imbalanced data can best enhance predictive accuracy?

This study aims to compare various strategies for managing imbalanced data sets, focusing specifically on oversampling, under sampling, and cost-sensitive approaches in terms of their impact on model performance, particularly looking at how they influence the accuracy and reliability of predictions in the context of loan default prediction. The goal is to identify which method not only achieves the best predictive accuracy but also maintains robustness without losing valuable information.

2.4 Scope

The focus of this thesis was to examine the impact of different supervised machine learning (ML) techniques on loan default prediction. This study specifically emphasized evaluating model performance through metrics such as precision, recall, F1-score, and the Area Under the Curve (AUC) score. The classifiers evaluated in this research included:

- Logistic Regression
- Decision Tree
- Random Forest

The Random Forest (RF) ensemble methods are recognized in existing literature for their superior performance compared to the Decision Tree (DT) and Logistic Regression (LR) models. However, the high level of interpretability and explainability associated with LR and DT results makes it essential to incorporate these models into this thesis. This inclusion allows for a comprehensive comparison and understanding of how each model contributes to the accuracy and usability of loan default predictions.

2.5 Related Works

- **Loan default prediction using machine learning models**

In the initial stages of statistical loan default prediction, the emphasis was primarily on linear classifiers such as Logistic Regression (LR), where predictors are combined linearly in the model. LR has been established as a benchmark in the field of credit risk analysis. This is particularly because the opaque nature of ensemble methods presents challenges in terms of interpretability, which is crucial for credit assessments (Dumitrescu et al. 2021). In linear models like LR, the importance of predictors in default prediction is indicated by the coefficients and their statistical significance, offering clear insights into determinants of default.

However, more complex ensemble methods, often described as a 'black box', lack this level of interpretability (Xia et al. 2021). Recently, non-linear approaches such as Decision Trees (DT), and Random Forests (RF) have been explored for default prediction. Unlike linear models, these tree-based models are capable of capturing non-linear relationships, discontinuities, and complex interactions between variables.

Additionally, tree-based methods such as DT and RF exhibit resilience to outliers in data and are invariant to monotonic transformations of predictors (Sigrist and Hirnschall 2019). They also handle issues of multicollinearity effectively, where linear models falter due to high correlations among predictors (Kruppa et al. 2013). Decision-tree-based classifiers further streamline the prediction process through automated iterations without the need for manual intervention, adding practical value to loan prediction tasks (Zhou et al. 2019).

These algorithms have shown promising results in tackling challenges posed by imbalanced and high-dimensional datasets. As such, tree-based algorithms have been recognized as advanced tools for loan default prediction in recent years. The advent of extreme gradient boosting, in particular, marks a significant development in the field, celebrated for its speed, efficiency, and the ability to run computations in parallel.

- **Imbalanced data**

In imbalanced classification challenges, two primary factors are minority interest and the rarity of instances. Minority interest highlights the importance of rare instances; for example, in fraud detection, the minority class, which is less frequent, is crucial. Rarity, meanwhile, indicates that instances of a specific class are few in relation to other classes. Imbalanced classification issues often stem from these two aspects combined, like in predicting rare diseases. Under such conditions, traditional machine learning algorithms, which are tailored to maximize accuracy, may falter. Although these models might achieve high overall accuracy, they often underperform in accurately predicting outcomes for the minority class.

Oversampling and undersampling are often used when talking about imbalanced datasets. Oversampling increases the representation of the minority class by replicating its instances, while undersampling decreases the majority class's instances. These methods are straightforward and can effectively even out class distributions, sometimes enhancing model performance (Chawla et al. 2002). However, oversampling may cause models to overfit, and undersampling can result in the loss of important data. The figure below illustrates how the over sampling and undersampling methods work:

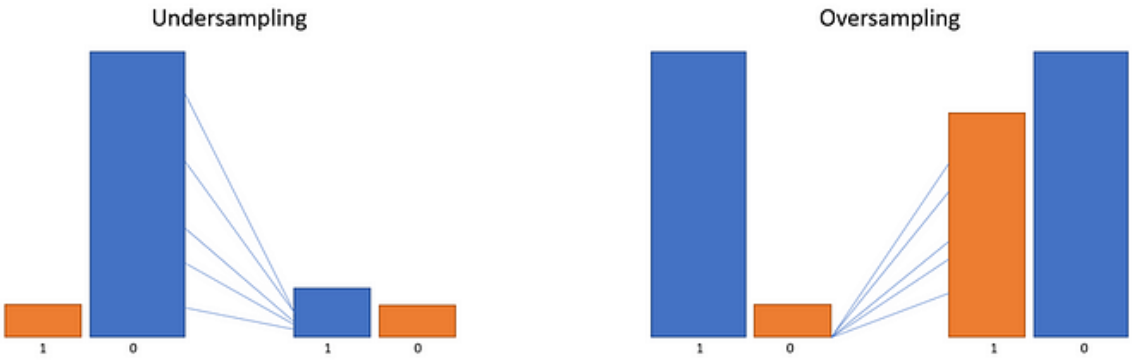


Figure 1 Oversampling and undersampling in rebalancing

To address these drawbacks, more advanced techniques like the Synthetic Minority Over-sampling Technique (SMOTE) have been introduced. SMOTE creates artificial instances of the minority class through interpolation among existing samples, thereby enriching the minority class without simple duplication (Chawla et al. 2002). This method aids in developing models that generalize better and has gained wide acceptance due to its effectiveness.

On the algorithmic level, solutions such as cost-sensitive learning and ensemble techniques have proven effective. Cost-sensitive learning adjusts the algorithm to place greater emphasis on the minority class by increasing the penalty for misclassifying it (Elkan 2001). Ensemble methods like the Balanced Random Forest (BRF) and Adaptive Boosting (AdaBoost) employ multiple models that handle different samples and aggregate their outputs to enhance performance on imbalanced data sets (Chen et al. 2004). These strategies are adept at managing complex data patterns and interactions, providing resilience against imbalances.

These varied techniques have shown considerable success in improving the management of imbalanced datasets in different fields. Consequently, they are essential for developing machine learning models that are accurate, reliable, and capable of better generalization and robustness in their predictions.

3. Methodology

3.1 Statistical Techniques

- Logistic regression:

Logistic regression serves as a classification mechanism that establishes boundaries between classes, determining class probabilities based on the distance from these boundaries. These probabilities are then categorized as 0 or 1, simplifying and refining the classification process. The regression equation is described as follows:

$$P(y = 1|X) = \frac{1}{1 + e^{-(a+b^T X)}}$$

Equation 1 Logistic Regression formula

In this equation, 'e' denotes Euler's number, while 'a' stands for an unknown parameter, and 'b' refers to a vector of unknown parameters. 'X' represents the predictor vector. The goal of the Logistic Regression (LR) model is to adjust the parameters 'a' and 'b' so that the equation,

shown as equation 1, produces an outcome as close to 1 as possible when identifying values that correctly fall into the positive class. LR is especially pertinent to this thesis because it provides detailed insights into the coefficients of the predictors and their statistical significance. This detailed analysis enables a specific evaluation of each predictor's influence on the likelihood of default, which presents a distinct advantage over ensemble methods that aggregate predictions.

Both DT and LR create decision boundaries to distinguish between classes. However, DTs are capable of dividing the decision space into smaller, non-linear regions with intricate boundaries, while LR generally uses a single linear boundary to divide the decision space (Kim 2016).

- Decision Tree:

In contrast to Logistic Regression (LR), Decision Trees (DTs) are nonlinear classifiers that categorize observations based on the responses to prior questions. A DT is structured like a tree: It starts with the root node at the base, followed by a series of decision nodes that represent possible choices, culminating in a leaf node that assigns the predicted class label (e.g., a defaulted loan). During the classification process, the Gini impurity metric is commonly used to identify the most effective predictor at each stage of tree construction. Gini impurity is a metric used in decision tree algorithms to assess how mixed a dataset is in terms of class distribution. It reflects the probability of incorrectly classifying a randomly chosen item from the dataset. A Gini impurity of 0 indicates a completely pure set, where all items belong to the same class, while higher values indicate greater diversity within the set. This metric is crucial for determining the optimal points to split the data during the construction of a decision tree

A DT evaluates splits across all predictors and selects the one that results in the greatest reduction in Gini impurity.

$$I_G = \sum_{i=0}^{1} p(i) * (1 - p(i))$$

Equation 2 GINI impurity's equation

In this context, IG represents the probability of misclassifying an observation, where 'I' denotes the class and p(i) is the probability of observing an observation of the i-th class.

A DT offers several advantages in loan classification scenarios. Firstly, the interpretation of results from a DT is straightforward. Once a DT is established, new observations can be quickly predicted using a series of if-then statements. This feature is particularly beneficial for loan default prediction as it allows DTs to reveal the relationships between predictors and the outcome. Secondly, as DTs are nonparametric, they generate if-then rules without any underlying assumptions about the distribution of the variables, which is advantageous given that loan data often exhibit nonlinear characteristics (Zhao and Zou 2021). Thirdly, DTs are adept at capturing the nonlinear relationships between variables, which is crucial for accurately modelling the complex dynamics of loan default prediction.

While introducing nonlinearity can complicate model interpretation, DTs manage to incorporate it in a way that remains understandable. Additionally, the risk of overfitting in DTs can be mitigated by reducing the tree's size, a method employed in this thesis. DTs are preferred for loan default prediction due to their intuitive and interpretable nature compared to more complex ensemble models. However, DTs are more sensitive to variations in data patterns than LR, meaning that minor changes in the training data can significantly alter the tree structure and impact predictions dramatically.

- Random Forrest:

Similar to DTs, RFs are nonlinear classifiers. RF is based on an ensemble technique involving many individual DTs, utilizing a method known as bagging. In this method, random subsets of the training data are created with replacement. For classification tasks, each DT independently predicts the output class, and the class receiving the majority of votes from these DTs is chosen as the final output. The procedure for the random forest algorithm is outlined as follows:

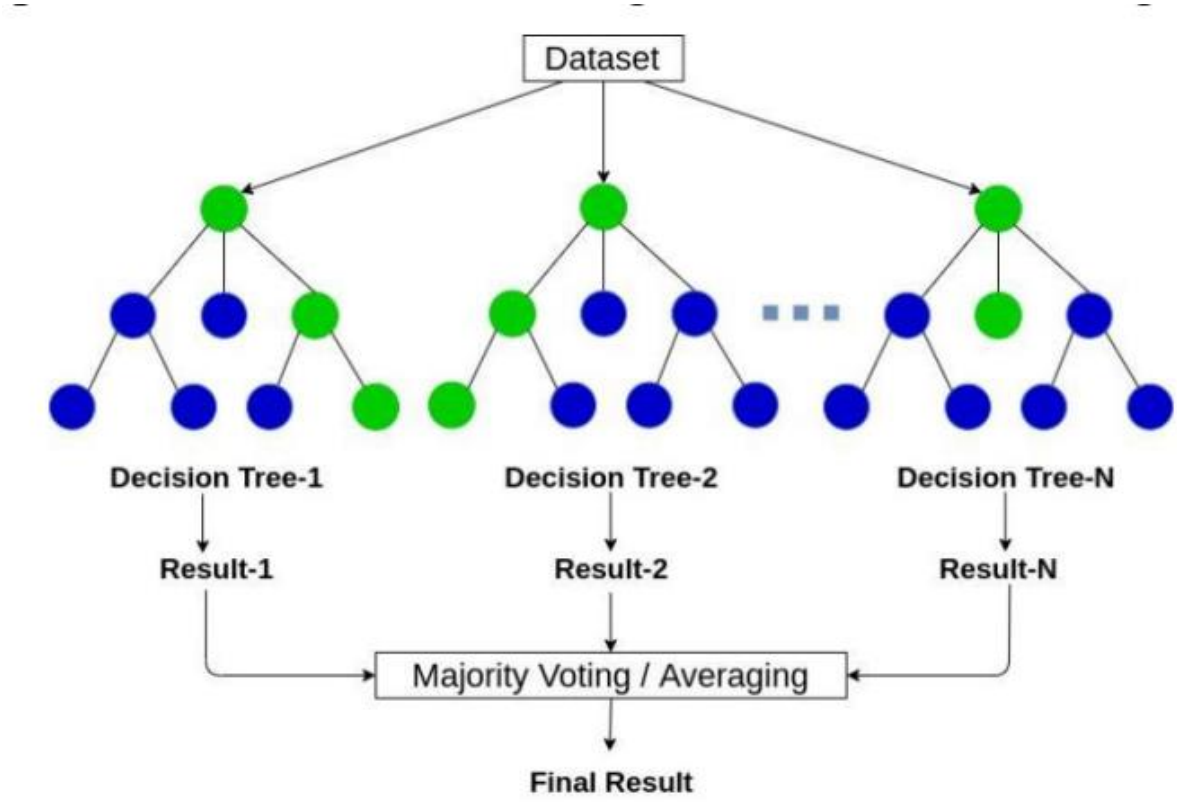


Figure 2 Diagram of random forest (Breiham, 2001)

The majority vote is calculated as follows:

$$\bar{f}(X) = sign(sign(\sum_{i=0}^{T} f_i(x)))$$

Equation 3 The calculation of majority vote in RF

Where ƒi represents the i-th decision tree in the ensemble, and "sign" denotes the signum function, which is defined as:

$$[0,\infty) \to \{0,1\}, sign(x) = 0 \Leftrightarrow x = 0$$

As mentioned above, each internal node of a DT evaluates the reduction in Gini impurity for each predictor, selecting the predictor that offers the most significant decrease in impurity for the node. In RFs, each tree is allowed to grow fully, which, despite the potential instability of individual trees, results in an aggregated classifier with reduced variance compared to single DTs. Thus, RFs are generally more resistant to overfitting than individual DTs. The random selection of predictors and bootstrapping contribute to the generation of uncorrelated trees in RFs. Adjusting the number of variables randomly sampled at each node significantly enhances the Area Under the Curve (AUC) metric (Probst, Wright and Boulesteix, 2019). In this optimization process, the number of variables at each split was adjusted on the validation set, using AUC (Area Under the Curve) as the evaluation metric, and employing a 10-fold cross-validation technique. The range for the number of variables considered spanned from 1 to 15. A key benefit of this ensemble method is its tendency to deliver superior predictive performance compared to using a single Decision Tree. This improvement arises because a group of weaker learners combines to form a more robust learner, typically allowing a Random Forest (RF) to surpass the performance of a standalone DT. Nonetheless, RF models are generally less interpretable than both Logistic Regression (LR) and DT models.

3.2 Performance Metrics

In this thesis, the terms "true positive" (TP) and "false positive" (FP) refer to non-performing loans (NPLs) that are correctly and incorrectly classified, respectively. "True negative" (TN) and "false negative" (FN) describe loans that are accurately identified as fully paid and loans that are erroneously labelled as defaults, respectively. Although accuracy is a common metric in binary classification, it can be misleading when applied to the imbalanced dataset used in this thesis. As a result, recall and precision are prioritized over accuracy in risk modelling. In particular, recall is more crucial than precision, as a false negative could lead to missed detection of potential defaults (Wang and Ni 2019). Additionally, the F1-score, which represents the harmonic mean of precision and recall, is emphasized as an essential measure in this analysis, offering a balanced perspective on model performance:

- Accuracy: The proportion of correctly classified cases to all cases in the set represents predictive accuracy. The higher this proportion, the better the model's performance.

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

Equation 4 Accuracy formula

- Precision: It is a metric that estimates the likelihood that a positively predicted value is accurate.

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

Equation 5 Precision formula

- Recall: Recall is the fraction of true positive cases that the model correctly identifies, typically calculated as the ratio of true positives to the total of true positives and false negatives. Essentially, it evaluates the model's capacity to identify every instance of the positive class without omitting any.

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

Equation 6 Recall equation

- F1-Score: The F1 score is derived from the harmonic mean of precision and recall, and optimizing the F1 score means enhancing both precision and recall at the same time.

$$F_1 - score = 2 * \frac{precision * recall}{precision + recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

Equation 7 F1-score formula

- Specificity: Specificity represents the proportion of true negatives that a model accurately predicts, generally determined by the number of true negatives divided by the sum of true negatives and false positives. It measures the model's ability to correctly recognize all instances of the negative class without mistakenly labelling any positive cases as negative.

$$Specificity = \frac{True\ negative}{True\ negative + false\ positive}$$

Equation 8 Specificity formula

Moreover, confusion matrix which includes details on predicted classifications (positive and negative) compared to the actual classifications (positive and negative). The confusion matrix, presented in Table 1, details measures such as accuracy, precision, recall, specificity, F1-score.

Table 1 Confusion Matrix

		Actual	
		Fully paid	Defaulted
Predicted	Fully paid	TN	FN
	Defaulted	FP	TP

3.3 Dealing with Imbalanced data

Addressing this imbalance is crucial before proceeding with the data analysis. If this step is overlooked, the analysis might be skewed due to the overwhelming presence of the majority

class. Such a scenario could lead to results that show high accuracy overall but are ineffective at detecting loan defaults, which is the primary focus of this thesis. This discrepancy necessitates techniques such as resampling or using advanced algorithms designed to enhance sensitivity towards the minority class to ensure more accurate and meaningful outcomes. Consequently, several methods for managing imbalanced data are detailed below:

3.3.1 SMOTE (Synthetic Minority Oversampling Technique):

An oversampling technique first introduced in 2002 by Nitesh V. Chawla involves creating new samples of the minority class through linear combinations.

In detail, the method can be explained as:

- A minority class instance xi is chosen as the root sample for generating new synthetic samples.
- The K nearest neighbours of xj are identified
- n of these K neighbours is randomly selected for interpolation.
- The difference between xj and the selected neighbours is calculated. Using this difference, n synthesized samples are created according to the following formula:

$$x_{synth}^{(j)} = x^{(j)} + gap^{(j)} \times [x_{neighbor}^{(j)} - x^{(j)}]$$

Equation 9 The equation of synthesized samples

Here, gap(j) represents a uniformly distributed random variable ranging from 0 to 1 for the j-th feature, and n specifies the quantity of oversampling required.

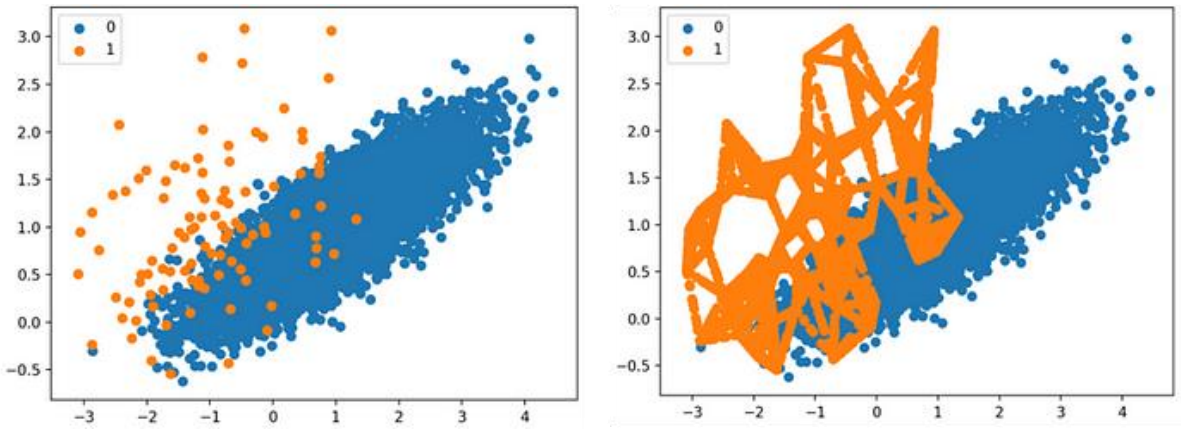


Figure 3 Example of dataset before and after applying SMOTE.

As a general guideline, it is imperative to divide the data into training and test sets before implementing any oversampling or undersampling techniques.

Applying oversample before splitting the data can lead the model to merely memorize specific data points, resulting in overfitting and poor performance when applied to the test data. Such data leakage can lead to the development of predictive models that are unrealistically optimistic or even entirely invalid.

3.3.2 SMOTE extensions: Tomek Link

Tomek Links is an undersampling approach aimed at enhancing the effectiveness of the undersampling procedure by targeting instances that may cause issues. The approach is as follows:

- Define a Tomek Link: A Tomek Link occurs between two instances, one from the majority class and one from the minority class, when they are the nearest neighbours to each other. Specifically, if instance A from the majority class and instance B from the minority class are closer to each other than to any other instances in their respective classes, they establish a Tomek Link.
- Remove Majority instances: After identifying Tomek Links, the instances from the majority class involved in these links are removed. This is because these majority instances are typically near the decision boundary and may overlap with the minority class, reducing their effectiveness in differentiating between classes.

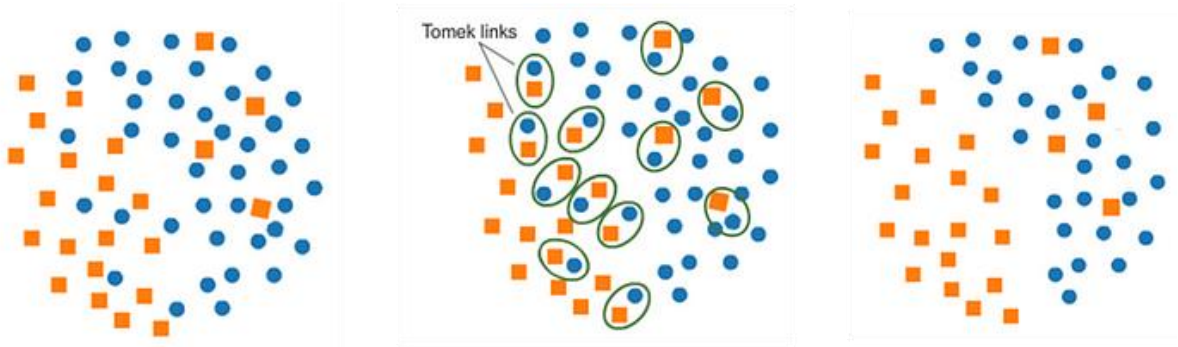


Figure 4 Illustration of Applying Tomek Links to an Imbalanced Dataset

Combining over-sampling of the minority (using SMOTE) class with under-sampling of the majority (TomekLink) class has been shown to enhance classifier performance compared to solely under-sampling the majority class.

3.3.3 Indirect Cost sensitive learning

This thesis also employs cost-sensitive learning to tackle the issue of an imbalanced dataset. A significant advantage of this approach is that it maintains the integrity of the original dataset, in contrast to other techniques like oversampling the minority class, which may introduce bias (Kim, Kwon, and Paik 2019). To implement indirect cost-sensitive learning, it is necessary to assign weights to the positive and negative classes within the training dataset. The precise weights designated for both classes are detailed in the equations presented below:

$$w_0 = \sqrt{\frac{\sum_{i=0}^{1} n_i}{n_0}} \quad w_1 = \sqrt{\frac{\sum_{i=0}^{1} n_i}{n_1}}$$

Equation 10 Cost sensitive equation

Here, 'ni' represents the number of observations of the i-th class in the training dataset, where i=0 indicates the negative class and i=1 denotes the positive class.

The weighted classes are used in the LR, DT and RF models.

In this dataset, it is evident that the target variable is significantly imbalanced with 88.3% being 0 (1,804,960 cases) and 11.7% being 1 (239,571 cases). Despite this imbalance, the weights assigned to each class are not drastically different: the majority class (0) has a weight of 1.064, while the minority class (1) has a weight of 2.921, approximately three times higher. This approach is known as balanced weighting. However, it should be noted that ideally, weights should be determined based on domain knowledge rather than a fixed ratio.

```
Weight of positive values 2.9213240030966183
Weight of negative values 1.0642975416480362
```

Figure 5 Weights for each class.

4. Findings and Analysis

4.1 Experimental Setup

4.1.1 Data Introduction

LendingClub, based in San Francisco, California, is an American peer-to-peer lending company. It was the pioneer in registering its offerings as securities with the Securities and Exchange Commission (SEC) and was the first to facilitate loan trading on a secondary market. LendingClub stands as the largest peer-to-peer lending platform globally.

The data is extracted from LendingClub dataset, it includes details about previous loan applicants and their default status. The objective is to uncover patterns that predict whether an individual is likely to default. This insight can be utilized for taking measures such as denying the loan, reducing the loan amount, or offering loans to higher-risk applicants at increased interest rates.

4.1.2 Data description

The dataset utilized in this thesis is publicly available on Kaggle and comprises historical data spanning from 2000 to 2007. The dataset initially contains over 2,500,000 entries across 27 variables. Following a detailed prescriptive analysis, it has been cleaned to include 2,044,531 unique loans, now represented by 19 variables. The target variable indicates whether a loan was paid in full or defaulted and is formatted as a dummy variable (0 = paid in full, 1 = defaulted). Out of the total, 239,571 loans were defaulted, while 1,804,960 loans were fully repaid, resulting in a default rate of about 11.7%

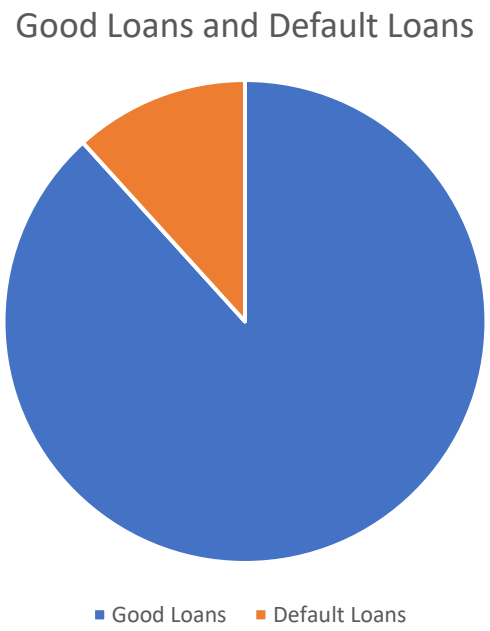


Figure 6 The portion of good loans and loan defaults

The detail description for each variables in the dataset can be found in table from Appendix C.

As can be seen from the Appendix C, it is readily apparent that the data primarily concentrates on individual loans. The original dataset included details such as the borrower's address, job title, and loan purpose, among others. However, these were presented in string format, which made them difficult to summarize and they contributed minimally to the final decision-making process, leading to their exclusion. For the remaining variables, they have been converted to numeric form to facilitate more effective analysis. Dummy variables have been implemented for these columns, with specifics as follows:

Table 2 The applied dummy variables

	Variable Name	Dummy Variable
1	Term	0 = 60 months, 1 = 36 months
2	Grade	6 = A, 5 = B, 4 = C, 3 = D, 2 = E, 1 = F , 0= G
3	Emp_length	2 = Over 10 years, 1 = Between 10 years and 1 year, 0 = Under 1 year
4	Home_ownership	4 = Own, 3 = Mortgage, 2 = Rent, 1 = Any + other, 0 = None
5	Loan_status	0 = Loan paid in full , 1 = defaulted

In order to have an overview of the dataset, the descriptive analysis of the loan variables is provided in Table 3:

Table 3 Descriptive Statistical summary of loan characteristics

Variable	Mean	Std	50%	Min	Max	Skew
Loan_amnt	15324.58	9226.01	13425	1000	40000	0.75
Term	0.70	0.46	1	0	1	-0.9
Int_rate	13.10	4.85	12.62	5.31	30.99	0.77
Installment	453.27	268.15	386.18	4.93	1719.83	0.98
Grade	4.33	1.26	4	0	6	-0.64
Emp_length	1.27	0.61	1	0	2	-0.22
Home_ownership	2.71	0.65	3	0	4	0.36
Annual_inc	80086.83	116345	67283.49	0	110000000	495.42
Verification_status	0.67	0.47	1	0	1	-0.73
Loan_status	0.12	0.32	0	0	1	2.38
Dti	18.70	11.44	17.87	-1	999	21.62
Open_acc	11.78	5.67	11	1	101	1.32
Pub_rec	0.2	0.57	0	0	86	11.92
Revl_bal	16941.2	23070.21	11563	0	2904836	12.86
Revol_util	50.68	24.57	50.7	0	892.3	0.01
Total_acc	24.37	12.02	23	2	176	1.01
Application_type	0.05	0.22	0	0	1	4.114
Mort_acc	1.55	1.9	1	0	94	1.8
Pub_rec_bankruptcies	0.13	0.36	0	0	12	3.46

The distribution of loan_status among categorical data can be shown below:

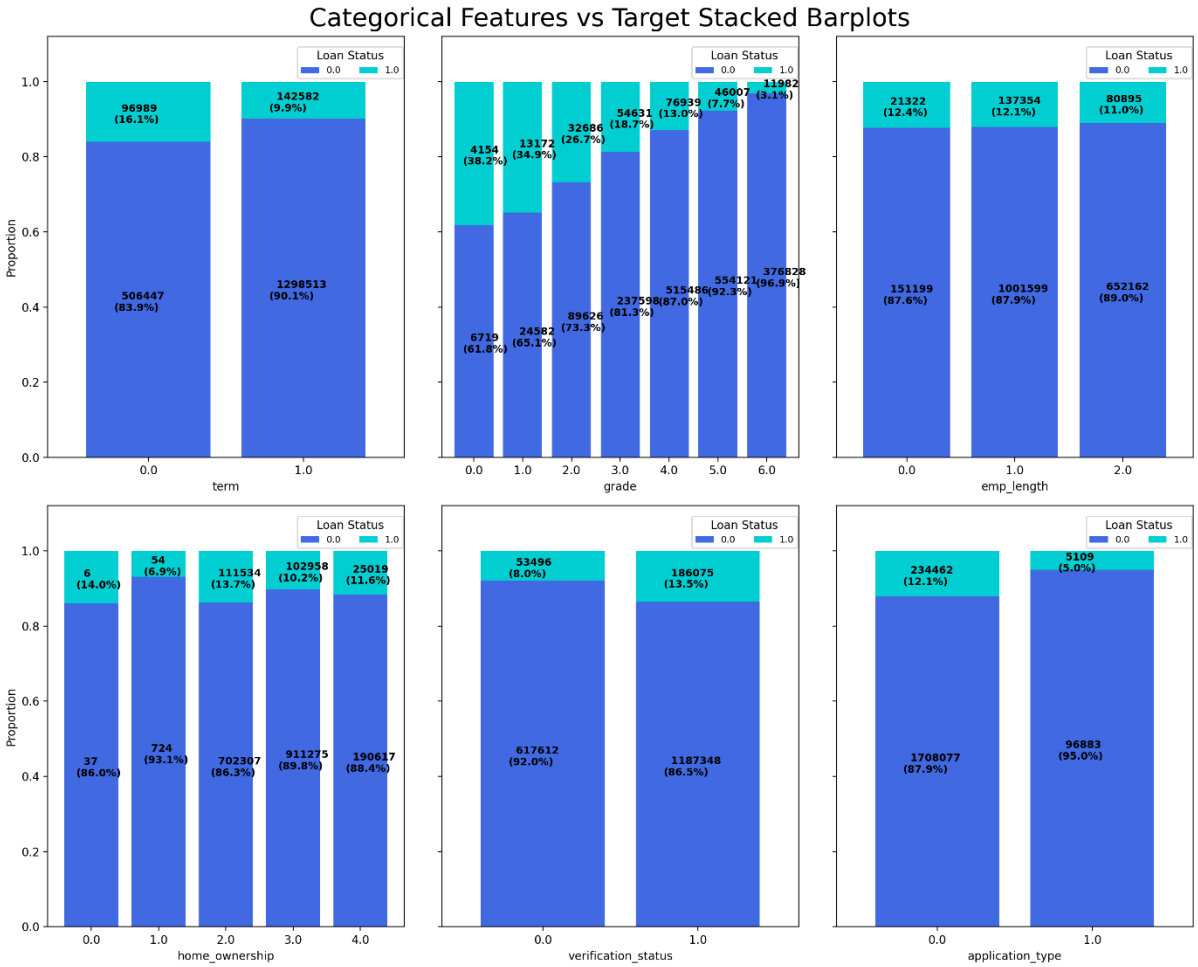


Figure 7 Visualization of categorical data

The distribution in emp_length, verification_status and application_type are almost equally. Loans with shorter terms (term = 0) exhibit a higher default rate (16.1%) compared to longer-term loans (term = 1) at 9.9%, indicating that shorter-term loans are riskier. The grade of the loan shows a strong inverse relationship with default rates, where lower grades (0, 1, 2) have higher default rates (38.2%, 34.9%, and 26.7% respectively), and higher grades (5 and 6) have much lower default rates (7.7% and 3.1% respectively). Employment length categories (0, 1, 2) present similar default rates around 12%, suggesting it is a less differentiating factor for default risk. Home ownership status impacts default rates variably; non-owners (category 0) have the highest default rate (14.0%), while other categories show mixed default rates ranging from 6.9% to 13.7%. Verification status indicates that loans with a verification status of 1 have a higher default rate (13.5%) compared to those without verification (8.0%). Lastly, application type significantly affects default risk, with type 0 applications having a 12.1% default rate and type 1 applications showing a lower rate of 5.0%. These findings highlight the importance of loan term, grade, verification status, and application type in predicting loan defaults, while employment length and home ownership provide more nuanced contributions to default risk.

Furthermore, the visualization of numerical data is presented below:

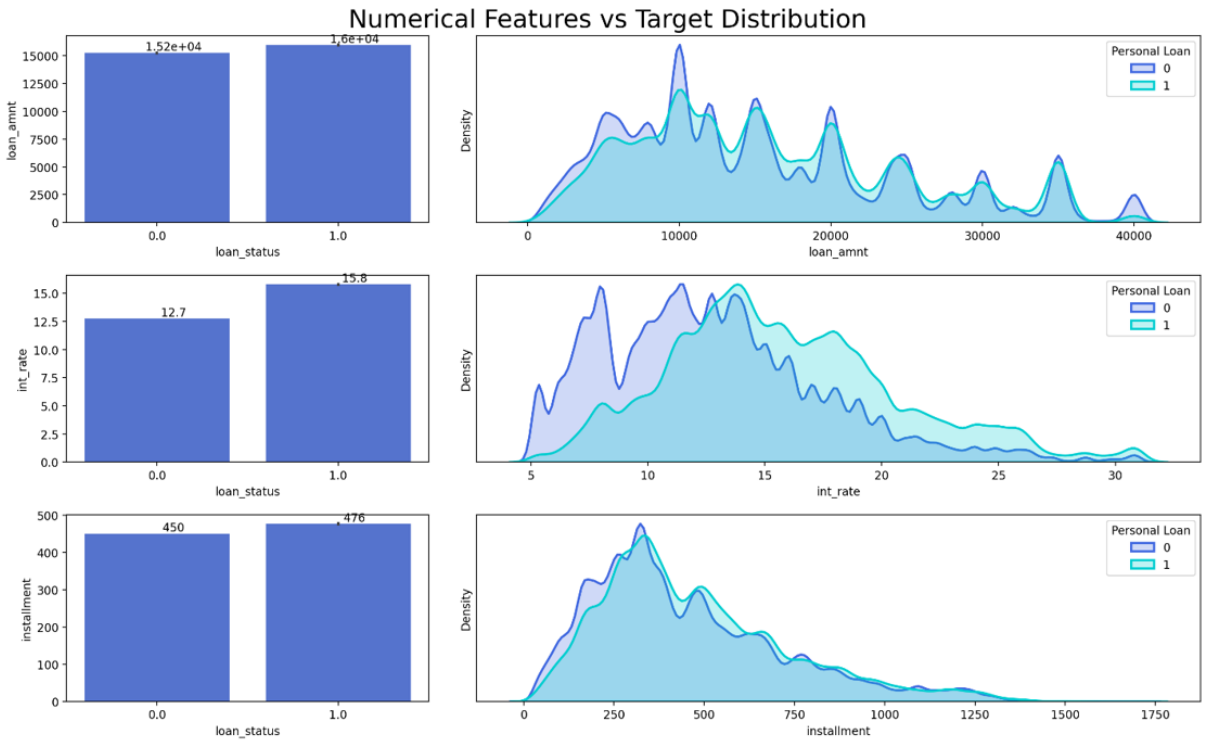


Figure 8 Visualization of Numerical data

The average loan amounts for both non-default and default statuses are similar, though defaults are slightly higher around the 10,000 to 20,000 (dollar). A notable difference is observed in interest rates, where defaulted loans have a higher average interest rate (15.8%) compared to non-defaulted loans (12.7%), indicating a strong correlation between higher interest rates and loan defaults. Additionally, while average instalment amounts are comparable between non-default and default loans (450 and 476, respectively), higher instalment values are more densely associated with defaults, suggesting an increased risk of default for loans with larger instalment payments.

4.2 Data Processing

The dataset is subjected to several preprocessing stages to eliminate redundant attributes, rescale observations, and manage extreme variable values, transforming them into a format that can be understood by the machine learning algorithm employed. These steps include:

- Data cleaning and reduction: outliers' detection and removing
- Feature standardization
- Feature normalization

As previously noted, all three models—LN, DT, and RF—require numerical inputs for training and prediction purposes. Consequently, some categories were transformed into new columns and assigned numerical values. However, three variables—home address, loan purpose, and borrower's job title—are also categorical. Encoding these would have led to a significant increase in the number of dimensions, potentially causing an increase in overfitting. Moreover, the job title variable does not provide more insightful information than annual income, which has a greater impact on the final decision. Therefore, these categorical variables were

excluded. Additionally, variables that were not relevant to the analysis, such as loan ID, city, and zip code, were also removed.

4.3 Exploratory Data Analysis

One goal of this thesis was to see how different models predict loan defaults. However, when predictors (variables used in the model) are correlated, it complicates understanding their individual effects on predicting defaults. This issue can make the model's estimates less precise and broaden the confidence intervals around these estimates. To address this, a correlation matrix (found in Appendix A) was examined to identify any strong correlations between predictors. For instance, there was a strong correlation between the loan amount and instalment variables. To avoid problems from these correlations (known as multicollinearity), only the loan amount was used for classification. Similarly, the opened account, public record of bankruptcies variables was dropped because it was strongly correlated with the total account variable.

4.4 Packages of Software

In this thesis, data manipulation and machine learning tasks are executed using Google Colab and the Python programming language. Additionally, SPSS software is also used to operate machine learning models, enhancing the analytical capabilities of the study. All the packages used in Colab are detailed in the table below:

Table 4 Google Colab packages

Package	Purpose
Numpy	Used for numerical operations on arrays and matrices.
Pandas	Data manipulation and cleaning
Lightbm	A gradient boosting framework that uses tree-based learning algorithms for machine learning tasks.
Sklearn	Machine learning technique
Matplotlib	Animated Visualization
Seaborn	Visualization
Scipy	Statistical tests
Tensorflow	Model training
Keras	Model training
Imblearn	Imbalanced dataset methods
Xgboost	Model training

4.5 Train test split

The dataset used for developing and evaluating the machine learning models was split into two parts, with 80% dedicated to training and 20% reserved for testing. The training portion was used to calibrate the models' parameters, and the testing portion was used for their assessment. For the Random Forest model, specifically, 10% of the training data was used to fine-tune the hyperparameters. There were 1,635,624 entries in the training set and 408,907

in the test set. To ensure uniform evaluation conditions, the same datasets were used for both training and testing all models. This uniformity was achieved by using a fixed seed before the data division. Cross-validation (CV) techniques were applied to assess the robustness of the models and prevent overfitting. Stratified CV was employed due to class imbalances, ensuring even representation of each class in every fold. Considering the large volume of data, a 10-fold CV method was utilized.

4.6 Results and analysis

4.6.1 Overall results of all three models

Considering the metrics of most interest in this analysis, RF exhibited the best predictive performance with regard to specificity, accuracy, and AUC under every handling imbalanced data method. However, while precision and F1-score are both excellent for cost-sensitive models, recall is relatively low, indicating that the positive class (loan default) is poorly predicted. Logistic Regression (LR) with oversampling displays balanced results with an accuracy, precision, recall, and F1-score of 0.66. Undersampling shows slight drops in precision, recall, and F1-score to 0.65, while cost-sensitive improves accuracy and precision significantly but sees a drop in recall to 0.57. Decision Tree (DT) with oversampling demonstrates high performance across all metrics, but undersampling leads to substantial drops, and cost-sensitive shows a lower accuracy and precision but improved specificity. Random Forest (RF) with oversampling achieves the highest accuracy (0.9301) and F1-score (0.9298), with nearly perfect specificity. However, undersampling results in significant performance drops, and cost-sensitive approaches show lower recall.

Overall, RF consistently shows the best performance, indicating it is the most reliable model in this context. Cost-sensitive LR and DT models improve specificity and accuracy but reduce recall, suggesting these models are better at correctly identifying non-defaulters but may miss actual defaulters. Undersampling methods generally reduce performance, indicating they might not be effective for this dataset. CSL improves precision and AUC for LR and DT but at the cost of recall, highlighting the trade-off between precision and recall. In this instance, CSL did not demonstrate effective performance across all the models tested. Its predictive capability was limited primarily to identifying majority cases, specifically those involving good loans. Unfortunately, CSL was notably less successful in accurately predicting instances of default loans with all the models.

In summary, RF with oversampling offers the best balance of high precision, recall, F1-score, and specificity, making it the most effective model for predicting loan defaults in this dataset, while cost-sensitive approaches require careful consideration based on specific application needs.

Table 5 Results of ML models

Model	Method	Accuracy	Precision	Recall	F1-Score	AUC	Specificity
LR	Oversampling	0.66	0.66	0.66	0.66	0.7199	0.6541
	Undersampling	0.65	0.65	0.65	0.65	0.7093	0.6569
	Cost-sensitive	0.8625	0.62	0.57	0.58	0.7080	0.9533
DT	Oversampling	0.8811	0.8812	0.8810	0.8810	0.88	0.8715
	Undersampling	0.5955	0.5955	0.5952	0.5949	0.5955	0.5983
	Cost-sensitive	0.8063	0.54	0.55	0.55	0.5470	0.8858
RF	Oversampling	0.9301	0.9373	0.9301	0.9298	0.9618	0.9946
	Undersampling	0.6632	0.62	0.739	0.674	0.7247	0.545
	Cost-sensitive	0.8823	0.66	0.5	0.48	0.7145	0.9981

4.6.2 Result of Logistic Regression

The results of the LR were analysed to gauge the significance and impact of individual contributions, as displayed in Table 6. The impact was measured using the estimated coefficient, and significance was evaluated based on the p-value. For continuous predictors, the estimated coefficient (βi) reflects the expected change in the odds of default for each unit increase in the predictor, changing by a factor of e^βi. For binary predictors, changing predictor j from the reference category 0 to 1 shifts the estimated odds of default by a factor of e^βj. The p-value was used to assess the statistical significance of each coefficient, with a significance threshold (α) set at 5%. Coefficients were considered statistically significant if their p-value was below this threshold. The intercept, set to assume a value of 0 for all predictors, was omitted from Table 7 as it did not present a realistic scenario within the framework of this thesis

Table 6 Variables in the equation of Logistic Regression

Variable	Estimate	SE	P-Value
Loan_amnt	.000	.000	<0.01
Term	-0.008	.003	0.008
Int_rate	-.074	.001	<0.01
Grade	-.838	.003	<0.01
Emp_length	-.095	.002	<0.01
Home_ownership	-.147	.002	<0.01
Annual_inc	.000	.000	<0.01
Verification_status	.311	.003	<0.01
Dti	.007	.000	<0.01
Pub_rec	.126	.002	<0.01
Revol_bal	.000	.000	<0.01
Revol_ulti	.004	.000	<0.01
Total_acc	.010	.000	<0.01
Application_type	-1.170	.007	<0.01

The result is taken from Logistic Regression model used oversampling method, which has stable outcomes from accuracy to specificity… All loan attributes were statistically significant, with the exception of the 'term' variable. Notably, Grade (-0.838), Application_type (-1.170), and Home_ownership (-0.147) exhibited negative coefficients, implying that higher grades,

specific application types, and owning a home decrease the likelihood of non-payment. On the other hand, Verification_status (0.311) and Pub_rec (0.126) showed positive coefficients, suggesting a greater probability of the event occurring when these elements are present. Other variables like Int_rate and Emp_length, while having smaller effect sizes, were still significant, underscoring their importance in the model. These findings indicate that both financial metrics and personal characteristics have a significant influence on the outcome, demonstrating that the model effectively captures the variance with accurate estimates.

Table 7 Result of LR with oversampling method

Classification Table[a]

			Predicted		
			loan_status		Percentage
	Observed		.00	1.00	Correct
Step 1	loan_status	.00	1177551	627409	65.2
		1.00	608726	1196234	66.3
	Overall Percentage				65.8

a. The cut value is .500

When compared the outcome with those from other methods, the findings were largely consistent in terms of which variables were statistically significant. The primary difference among imbalanced data methods, however, lies in the coefficients assigned to these variables. These differences in coefficients highlight how each method quantifies the impact of variables differently, suggesting variations in how they interpret the underlying data relationships and their influence on the predicted outcomes.

4.6.3 Result of Decision Tree

As mentioned above, two advantages of a DT are the interpretability and rapid manner in which new observations can be classified. In this model, visualizing all the splits remains difficult because decision trees often have many splits. A common issue with these trees is overfitting, which can be managed by pruning the trees using an optimal complexity parameter (cp). The cp determines the threshold where further splits don't significantly improve the model's fit. The default cp value is 0.01, but for complex scenarios, it's beneficial to adjust this value to a more lenient setting.

The optimal number of splits determined was three, as detailed in Appendix D. This finding is based on the results from the Oversampling method, which provided the best outcomes compared to other methods.

As illustrated, the variables home_ownership, grade, and emp_length were crucial for splitting the nodes. Each node included information about the class (positive or negative), the probability of default.

At the root node, the overall probability of loan default was established at 50%. The root node posed a question regarding whether the borrower's grade was above 6. If the answer was yes, processing moved to the right child node. This node revealed that loans with terms greater

than 6 had a much lower default probability of only 4.4%. This pattern was consistent in other nodes as well, where lower grades were associated with a higher default probability.

Following the grade, the home_ownership variable was used for further splits. It was observed that customers without a house or mortgage were more likely to default. The final split employed the emp_length variable to gauge the likelihood of loan default. Predictably, it was found that the longer someone's employment duration, the lower their risk of defaulting on a loan.

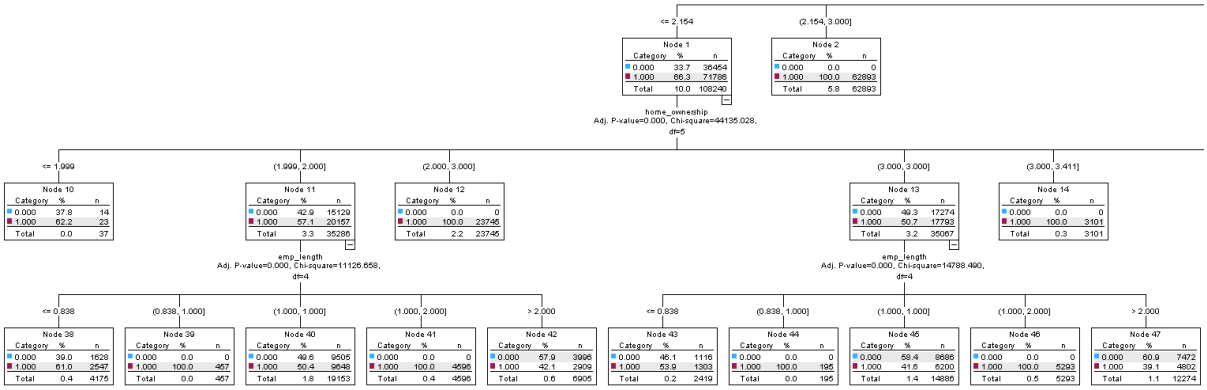


Figure 9 Graphical depiction of pruned DT

When compared to other models, regardless of the lower outcomes, the decision-making process for splitting nodes shows greater variation at each step. For instance, in the Random Forest (RF) model employing the undersampling method, the variable 'grade' was initially used for the first split. Subsequently, the variable chosen for further splits varied depending on the node. At the node corresponding to grade 3, the interest rate was used for further division, while at the node for grade 4, the application type was employed.

This pattern was consistent across other nodes as well. Generally, this indicates that the model was better at capturing the complex patterns within the dataset. However, the results were not entirely satisfactory as the model struggled to predict negative cases effectively, though it did reasonably well in predicting positive cases. This highlights the model's nuanced ability to handle diverse aspects of the dataset but also points to its limitations in achieving balanced predictive performance.

Table 8 Result of RF with undersampling method

Classification

Observed	Predicted .00	Predicted 1.00	Percent Correct
.00	39037	32565	54.5%
1.00	18782	53214	73.9%
Overall Percentage	40.3%	59.7%	64.2%

Growing Method: CHAID
Dependent Variable: loan_status
Test sample results are displayed.

A similar situation was observed with the indirect cost-sensitive method, which utilized a variety of variables to split nodes, including application_type, annual_inc, term, total_acc, among others. However, despite the diverse range of variables used, the overall predictive performance of this method was disappointing.

4.6.4 Result of Random Forrest

According to the data from Table 5, the oversampling method demonstrated superior performance compared to the other two methods, achieving 93% in accuracy, 93% in precision, 93.01% in recall, and 99.46% in specificity. This indicates an excellent balance in predicting both negative and positive cases. On the other hand, the undersampling and cost-sensitive methods did not yield as favorable outcomes. While the cost-sensitive learning (CSL) method achieved decent accuracy at 88.23% and an impressive specificity of 99.81%, the undersampling method performed poorly across all metrics, with only 66.32% accuracy and 54.5% specificity. Considering the Area Under the Curve (AUC), the oversampling method still ranked highest with an AUC score of 0.9618, followed by the undersampling method at 0.7247 and the cost-sensitive method at 0.7145.

4.6.5 Analysis

The traditional method of model evaluation relies on performance metrics, which we have gathered for classification purposes, including ROC, AUC, F1-Score, Accuracy, and Specificity. In this section, we will compare all the results to aid in comparing and selecting the best model. In the heatmap depicted in Figure 9, the colour coding was used to emphasize the metrics: red indicates the least desirable outcomes, while green signifies the most desirable outcomes.

Table 9 Heat map for model comparison

Model	Method	Accuracy	Precision	Recall	F1-Score	AUC	Specificity
LR	Oversampling	0.66	0.66	0.66	0.66	0.7199	0.6541
	Undersampling	0.65	0.65	0.65	0.65	0.7093	0.6569
	Cost-sensitive	0.8625	0.62	0.57	0.58	0.708	0.9533
DT	Oversampling	0.8811	0.8812	0.881	0.881	0.88	0.8715
	Undersampling	0.5955	0.5955	0.5952	0.5949	0.5955	0.5983
	Cost-sensitive	0.8063	0.54	0.55	0.55	0.547	0.8858
RF	Oversampling	0.9301	0.9373	0.9301	0.9298	0.9618	0.9946
	Undersampling	0.6632	0.62	0.739	0.674	0.7247	0.545
	Cost-sensitive	0.8823	0.66	0.5	0.48	0.7145	0.9981

Regarding performance metrics, the Random Forest model using the Oversampling method exhibited the best overall performance, while, surprisingly, the Decision Tree model with the Undersampling method showed the worst. The highest accuracy was achieved by the Random Forest model with the Oversampling method. Moreover, this model also proved to be the most suitable for our scenario of imbalanced data, achieving the highest accuracy and specificity.

The differences in performance metrics among the LR, DT, and RF classifications align with previous research findings. Studies by Kruppa et al. (2013) and Zhou et al. (2019) demonstrate that tree-based algorithms generally surpass the benchmark LR model in predicting loan defaults, consistent with the results presented in this thesis. Additionally, Malekipirbazari and Aksakalli (2015) highlighted that a cost-sensitive approach to addressing class imbalance positively affects the accurate prediction of non-performing loans (NPLs), although it also results in more incorrect predictions of fully paid loans. Furthermore, it is noted that in cases of highly imbalanced datasets, the undersampling method may not be effective as it could eliminate important patterns, thereby hindering effective model training. In contrast, the oversampling method proved to be a good fit, as it was able to capture and learn patterns effectively.

Furthermore, one of reason for CSL's underperformance may be attributed to the method used to determine the cost values. In this study, a basic cost structure was applied, which may not have accurately reflected the true cost dynamics associated with misclassifying loans.

Research has shown that the effectiveness of CSL depends heavily on accurate cost estimation and calibration (Elkan, 2001). If the costs, particularly those associated with false positives (i.e., failing to predict a default), are underestimated, the model may still favor the majority class, thereby reducing its ability to correctly identify high-risk loans. Ideally, these cost values should be informed by domain-specific knowledge, such as the financial impact of loan defaults on the lending institution, to ensure they accurately represent the real-world consequences of misclassification (Elkan, 2001).

Another factor affecting CSL's performance could be its interaction with the chosen machine learning models. For example, models like Random Forest, while generally powerful, may not fully capitalize on the benefits of CSL without additional tuning. The impact of CSL might be diluted when used with ensemble methods like Random Forest, where multiple weak learners contribute to the final prediction. If the cost sensitivity is not consistently enforced across all learners within the ensemble, the overall effect can be less pronounced, leading to suboptimal performance (Chen, Liaw, & Breiman, 2004). Ensuring that CSL is integrated effectively across all components of an ensemble model is crucial for maintaining its intended impact on predictive accuracy.

5. Conclusion

The key findings of this study demonstrate that the Random Forest model, particularly when paired with oversampling techniques, outperforms other machine learning models such as Logistic Regression and Decision Trees in predicting loan defaults. The Random Forest model achieved the highest accuracy, precision, recall, and F1-Score, indicating its superior ability to balance and correctly predict both default and non-default cases. The study also found that handling imbalanced data is crucial for improving predictive accuracy, with oversampling showing the most consistent benefits across different models. Cost-Sensitive Learning (CSL) provided some improvements in accuracy and specificity, but it did not perform as well in recall and F1-Score, highlighting the challenges in applying CSL without precise cost calibration.

The research questions posed in this thesis are restated and addressed below:

Research question 1:

'Which machine learning algorithm from the selected group shows the best results in predicting loan defaults, according to specific model evaluation criteria?'

In conclusion, the Random Forest model demonstrated superior predictive capabilities for loan default prediction, outperforming both the Logistic Regression (LR) and Decision Tree (DT) models, as corroborated by existing literature. This superior performance is likely due to Random Forest's ability to handle large datasets with higher dimensionality and its robustness against overfitting. Additionally, the Decision Tree model showed better performance than the Logistic Regression in certain aspects, aligning with findings from previous studies.

Research question 2:

'How much the ensemble method can improve the accuracy of predicting loan defaults?'

The ensemble method had been proven that can significantly enhance the accuracy of predicting loan defaults by combining the strengths of multiple machine learning models. Techniques such as Bagging, Boosting, and Stacking integrate the outputs of several base learners to produce a more robust and accurate prediction. In this thesis, Random Forests, which use Bagging, could improve accuracy by aggregating the predictions of multiple decision trees, reducing variance and avoiding overfitting. Results shows that Compared to LR and DT, RF with any method shows a 5% to 30% improvement in accuracy, recall, and F1-Score, underscoring RF's superior capability in effectively predicting loan defaults across various data handling techniques. These methods are particularly effective in handling complex data structures and interactions, leading to more reliable and precise loan default predictions.

Research question 3:

"Which method for managing imbalanced data can best enhance predictive accuracy?

Among various techniques for managing imbalanced data, the Synthetic Minority Over-sampling Technique (SMOTE) has proven to be the most effective in enhancing predictive accuracy. SMOTE works by generating synthetic samples for the minority class, thus balancing the dataset without simply replicating existing samples. This method helps the model learn the underlying patterns of the minority class more effectively. Studies have demonstrated that SMOTE, when combined with robust classifiers like Random Forest or ensemble methods, significantly improves the model's ability to predict minority class instances, resulting in better overall predictive performance.

However, it is important to note that in scenarios where the weights for the Cost-Sensitive Learning (CSL) are calculated more accurately, the performance of the model could improve. Generally, the Synthetic Minority Over-sampling Technique (SMOTE) performs well without the need for additional calculations.

References

Breiman, L., 2001. Random forests. *Machine Learning*, 45(1), pp.5-32.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). 'SMOTE: Synthetic Minority Over-sampling Technique'. Journal of Artificial Intelligence Research, 16, pp. 321-357.

Chen, C., Liaw, A., & Breiman, L. (2004). 'Using Random Forest to Learn Imbalanced Data'. University of California, Berkeley.

Dumitrescu, E-I., Hué, S., Hurlin, C., et al. (2021). 'Machine learning or econometrics for credit scoring: Let's get the best of both worlds'.

Elizalde, A. (2005). 'Do we need to worry about credit risk correlation?' The Journal of Fixed Income, 15(3), pp. 42–59.

Elkan, C. (2001). 'The Foundations of Cost-Sensitive Learning'. In Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2 (IJCAI'01), B. Nebel (Ed.), Vol. 2. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 973–978.

Ereiz, Z. (2019). 'Predicting default loans using machine learning (optiml)'. In 2019 27th Telecommunications Forum (TELFOR), pp. 1–4, IEEE.

Fitch Ratings, 2024. *Fitch Reports U.S. Loan Default Rate at 3.04% in 2023, Forecasts Further Rise*. [online] Available at: https://www.lsta.org/news/fitch-us-loan-default-2023/ [Accessed 13 August 2024].

Ghosh, A. (2017). 'Sector-specific analysis of non-performing loans in the US banking system and their macroeconomic impact'. Journal of Economics and Business, 93, pp. 29–45.

He, H., & Ma, Y. (2013). 'Imbalanced Learning: Foundations, Algorithms, and Applications'. Wiley-IEEE Press.

Hodge, V. J., & Austin, J. (2004). 'A Survey of Outlier Detection Methodologies'. Artificial Intelligence Review, 22(2), pp. 85-126.

Hussain, A., Khalil, A., & Nawaz, M. (2013). 'Macroeconomic determinants of non-performing loans (NPL): Evidence from Pakistan'. Pakistan Journal of Humanities and Social Sciences, 1(2), pp. 59–72.

Kim, K. (2016). 'A hybrid classification algorithm by subspace partitioning through semi-supervised decision tree'. Pattern Recognition, 60, pp. 157–163.

Kim, Y., Kwon, Y., & Paik, M. C. (2019). 'Valid oversampling schemes to handle imbalance'. Pattern Recognition Letters, 125, pp. 661–667.

Klein, N. (2013). 'Non-performing loans in CESEE: Determinants and impact on macroeconomic performance'. International Monetary Fund.

Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). 'Consumer credit risk: Individual probability estimates using machine learning'. Expert Systems with Applications, 40(13), pp. 5125–5131.

Lee, J., & Rosenkranz, P. (2019). 'Nonperforming Loans in Asia: Determinants and Macrofinancial Linkages'. ADB Economics Working Paper Series. No. 574. Manila: Asian Development Bank.

Lee, Y., & Poon, S-H. (2014). 'Forecasting and decomposition of portfolio credit risk using macroeconomic and frailty factors'. Journal of Economic Dynamics and Control, 41, pp. 69–92.

LSTA, 2023. *Loans Close 2023 With Second Highest Return on Record*. [online] Available at: https://www.lsta.org/news/loans-2023-second-highest-return/ [Accessed 13 August 2024].

Malekipirbazari, M., & Aksakalli, V. (2015). 'Risk assessment in social lending via random forests'. Expert Systems with Applications, 42(10), pp. 4621–4631.

Sigrist, F., & Hirnschall, C. (2019). 'Grabit: Gradient tree-boosted tobit models for default prediction'. Journal of Banking & Finance, 102, pp. 177–192.

Sundaramahadevan, V. (2021, August). 'Credit EDA Case Study'. Retrieved from Kaggle: https://www.kaggle.com/venkatasubramanian/credit-eda-case-study

S&P Global Ratings, 2023. The U.S. Leveraged Loan Default Rate Could Reach 2.5% By December 2023. [online] Available at:

https://www.spglobal.com/ratings/en/research/articles/2023-us-leveraged-loan-default-rate-forecast [Accessed 13 August 2024].

Wang, Y., & Ni, X. S. (2019). 'A xgboost risk model via feature selection and bayesian hyper-parameter optimization'. arXiv preprint arXiv:1901.08433.

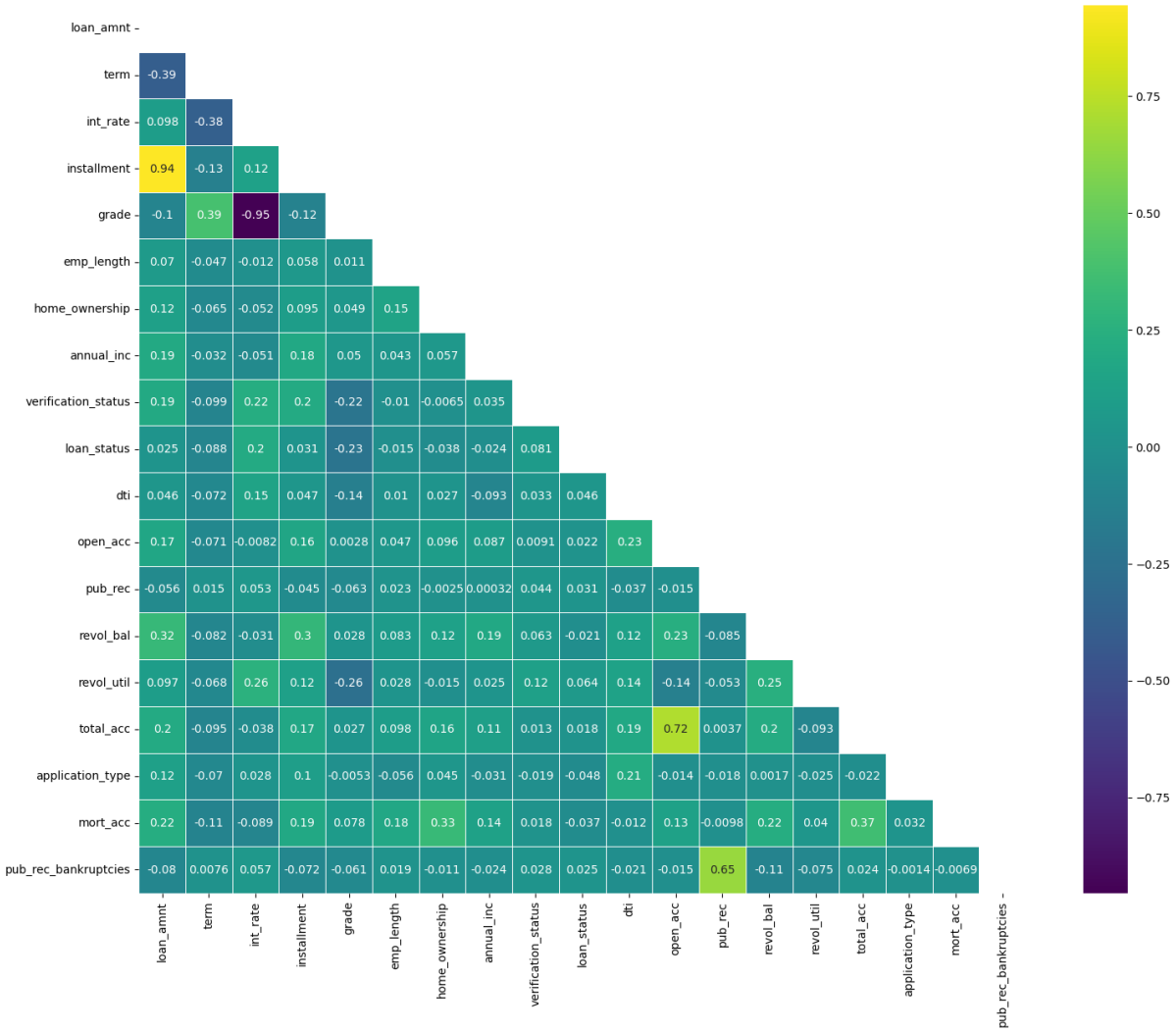
Xia, Y., Li, Y., He, L., Xu, Y., & Meng, Y. (2021). 'Incorporating multilevel macroeconomic variables into credit scoring for online consumer lending'. Electronic Commerce Research and Applications, 49, 101095.

Zhao, S., & Zou, J. (2021). 'Predicting loan defaults using logistic regression'. Journal of Student Research, 10(1).

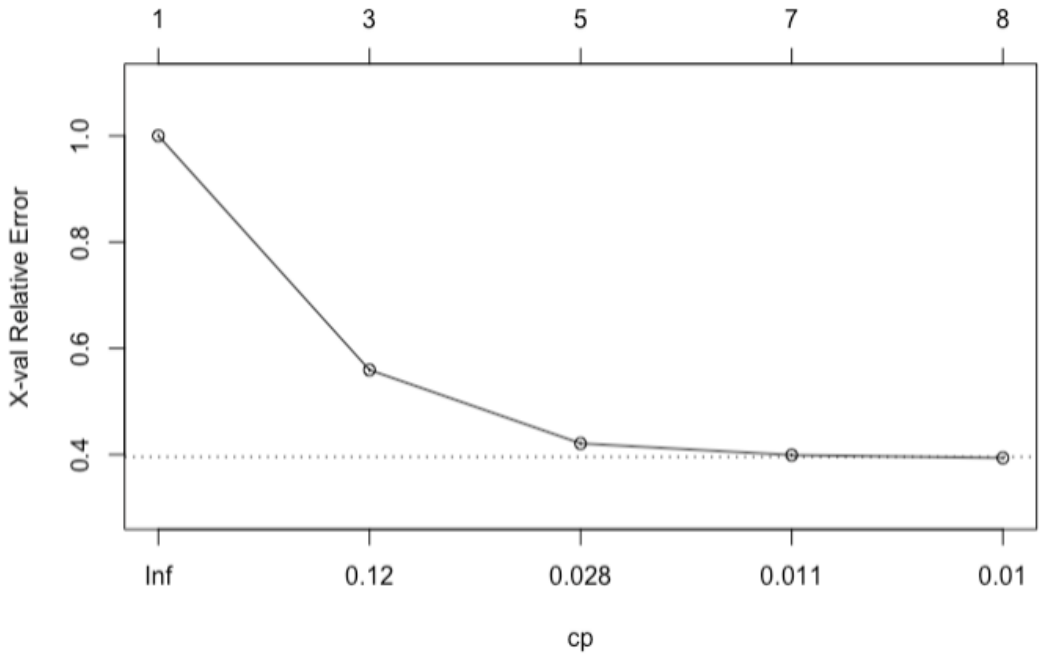
Zhou, J., Li, W., Wang, J., Ding, S., & Xia, C. (2019). 'Default prediction in p2p lending from high-dimensional data based on machine learning'. Physica A: Statistical Mechanics and itsApplications, 534, 122370.

Appendices

Appendix A: The correlation matrix for loan and macroeconomic variables shows that, with some exceptions like disbursement gross and approved amounts, correlations among predictors are typically low



Appendix B: Complexity plot for a decision tree, where the lower horizontal axis displays the level of complexity, the vertical axis indicates the error rate, and the upper horizontal axis represents the number of splits.



Appendix C: The description of variable in the dataset

	Variable Name	Description
1	Loan_amnt	The amount listed is the loan amount initially applied for by the borrower. If the credit department decides to reduce the loan amount at any time, this change will be reflected in the listed value.
2	Term	The number of payments on the loan, expressed in months. The terms available are either 36 or 60 months.
3	Int_rate	The interest rate applied to the loan.
4	Installment	The monthly payment due from the borrower if the loan is issued.
5	Grade	The loan grade assigned by Lending Club
6	Emp_length	The duration of the borrower's employment, ranging from 0 to 10 years, where 0 represents less than one year and 10 represents ten or more years.
7	Home_ownership	The borrower's home ownership status as stated during registration or as obtained from the credit report. Possible values include RENT, OWN, MORTGAGE, and OTHER.
8	Annual_inc	The borrower's self-reported annual income during registration.
9	Verification_status	Indicates whether the borrower's income was verified by Lending Club, not verified, or if the income source was verified.
10	Loan_status	The current status of the loan.
11	Dti	A ratio calculated using the borrower’s total monthly debt payments on all debt obligations (excluding mortgage and the requested Lending Club loan) divided by the borrower’s self-reported monthly income.
12	Open_acc	The number of open credit lines in the borrower’s credit file.
13	Pub_rec	The number of derogatory public records.
14	Revl_bal	The total balance of revolving credit.
15	Revol_util	The amount of credit the borrower is using relative to all available revolving credit.
16	Total_acc	The total number of credit lines currently in the borrower's credit file.
17	Application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers.
18	Mort_acc	The number of mortgage accounts.
19	Pub_rec_bankruptcies	The number of public record bankruptcies.