# Advanced Stacking Ensemble Learning for Robust Loan Default Risk Detection


Submitted By

**Mushfika Rahman Mridu**
**(202-35-3114)**
**Department of Software Engineering**

Supervised By

**Mr. Md Khaled Sohel**
**Assistant Professor**
**Department of Software Engineering**

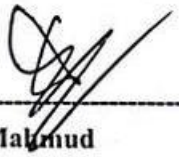

A thesis submitted as a partial requirement for the completion of the Bachelor of Science degree in Software Engineering.

Fall 2025

## APPROVAL

This thesis titled on "**Advanced Stacking Ensemble Learning for Robust Loan Default Risk Detection**", submitted by **Mushfika Rahman Mridu (ID: 202-35-3114)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

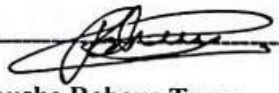
### BOARD OF EXAMINERS



---

**Chairman**

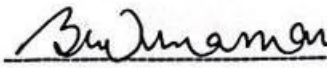
**Dr. S M Hasan Mahmud**
**Associate Professor**
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University



---

**Internal Examiner 1**

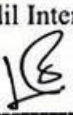
**Tapushe Rabaya Toma**
**Assistant Professor**
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University



---

**Internal Examiner 2**

**Khalid Been Badruzzaman Biplob**
**Lecturer (Senior Scale)**
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University



---

**External Examiner**

**Dr. Md. Sazzadur Rahman**
**Professor**
Institute of Information Technology
Jahangirnagar University

# DECLERATION

I hereby declare that the work presented in this thesis is my own original research work, carried out under the supervision of **Mr. Md Khaled Sohel,** Assistant Professor, Department of Software Engineering, Daffodil International University.

This work has not been submitted anywhere, either in whole or in part, for any degree, diploma, or publication in this or any other university. All sources of information used in this thesis have been duly acknowledged.

**Supervised By**

Mr. Md Khaled Sohel
Assistant Professor
Department of Software Engineering

**Submitted By**

Mushfika

Mushfika Rahman Mridu
ID: 202-35-3114
Department of Software Engineering

# ACKNOWLEDGEMENT

To start with, I would like to know that I would like to thank Almighty Allah to render me that strength, endurance, and determination that I needed to complete my study project. It is through His blessing that this would not have been possible. I would like to express my deepest thanks to my excellent parents who have always provided encouragement and support as well as guidance in my academic years. I have always been impressed by their sacrifices and trust in my abilities. I would like to express my deep sense of appreciation to **Prof. Dr. Imran Mahmud** who was a renowned Head of the Department of Software Engineering, and who offered the facilities and a favorable academic atmosphere that played a significant role in the achievement of the given endeavor. I owe a debt of gratitude to my supervisor, **Mr. Md Khaled Sohel**, who gave me the much-needed advice, the perseverance, and the incisive criticism throughout the entire research process. His expertise and assistance have greatly influenced this thesis. Also, I would like to thank the whole faculty of the Department of Software Engineering due to their dedication to academic excellence, supplying me with the information and skills that enabled me to achieve a successful completion of my project.

Finally, I would like to briefly note that my friends, classmates, and staff at Daffodil International University (DIU) helped and cooperated with me during this attempt, supported me, engaged in productive conversations, and offered encouragement.

# ABSTRACT

This dissertation suggests an entirely based machine learning solution, which is expected to significantly enhance loan default assessmetnt. We conducted a close comparison analysis of various machine learning and deep learning framework with a sharp instrument of XGBoost to boost the entire forecasting framework.

This will be enhancedt through finding a better pre-processing methods, like outlier management by winsorization and data normalization with resilience scaling. Multiple resampling methods have been hard investigated; we identified the hybrid SMOTE + ENN as the most effective option with respect to balancing unbalanced datasets, achieving an impressive 90.49 percent accuracy, 94.61 percent precision and 92.02 percent recall. We discovered 48 optimal predictors via Recursive Feature Elimination with Cross-Validation (RFECV), with interest rate, FICO score and loan term being the most significant ones. It will be foundeed upon our novel stacking ensemble model, that cleverly involves a multitude of base learners' predictions. This ensemble showed outstanding performance of 93.69 percent accuracy rate, 95.59 percent preciseness rate, 95.55 percent recall rate, and 97.81 percent Area Under the Receiver Operating Characteristic Curve (AUC), which is significantly high compared to individual model rates. Moreover, SHapley Additive exPlanations (SHAPs) are very transparent, and their components lend to practical understanding of the factors behind default prediction. This powerful, understandable, and movable paradigm offers financial organizations with a powerful tool to manage danger, reduce casualties, and increase lending choices in diverse datasets.

**Keywords**: credit default prediction; ensemble learning; machine learning; financial risk modeling; explainable AI; data imbalance.

# TABLE OF CONTENTS

# Chapter 1

# Introduction

## 1.1 Context and Relevance

Anticipating loan defaults is a crucial requirement of financial institutions like banks, credit bureaus and lending community. It is critical that these institutions analyze the availability of risk on offering money to the likely borrowers. With the assistance of accurate default prediction models, lenders are capable of minimizing costly lending, making efficient lending choices, and guaranteeing long-term stability and profitability. The necessity of robust credit risk assessment instruments fell within the context of the numerous defaults that took place during the 2008 global financial downturn that resulted in extreme economic turmoil and decline of confidence in financial institutions among people [1] [2].

Traditionally, financial companies analyzed the loan application based on the usual credit scores such as logistic rests and decision trees. However, these models tend to be associated with the issues of successful predicting of defaults especially when large and complex datasets are used, which involve a variety of features. As an illustration, non-linear associations amid features which play a substantial part in loan default prediction are often disregarded by the conventional models. The necessity of accuracy and flexibility of prediction instruments causes the expanding need to make use of new and improved fore-telling means as the new landscape of data gathering and fiscal danger has undergone changes. Machine learning (ML) can become one of the possible solutions to these challenges. The use of ML algorithms has the ability to review the large volumes of data and reveal complex patterns that may go undetected by the conventional models. In particular, ensemble and stacking, boosting and random forests have been found to be very effective in enhancing the forecast of underlying default models. Ensemble methods can use the characteristics of each model and reduce the risk of overfitting which is common to single models trained on complex data, by combining a significant number of base models [3][4].

Moreover, the issue of imbalance of the classes is among five principal issues on the issue of loan default prediction. Due to the large proportion of non-defaulters compared to the defaulters in normal data of loans, the skewed models are used to predict the majority class.

The imprecision in the classification may result in models that seem to be precise overall but fail to identify defaulters in appropriate ways. This challenge has to be addressed in order to come up with a reliable algorithm capable of identifying high-risk borrowers. This study was driven by the need to have more accurate, scaled and reliable solutions to the prediction of loan defaults. To improve prediction accuracy, particularly with the minority category, (i.e. loan defaulters) this work aims at developing a stacking ensemble method, which combines many machine learning algorithms. To improve the rate of performance of the model, the paper also considers alternative methods to control the issue of class imbalance, feature selection, as well as hyperparameter optimization [5].

## 1.2 Research Gap and Challenges

In loaning money, the financial institutions are very vulnerable to risks and estimation of the probability of loan defaults is one of the most crucial measures of minimising these risks. The problem of the class imbalance on which the largest part of the data was nondefaulters remains a major limitation toward producing an efficient prediction model, despite the innovation of machine learning methods. Natural applicability of most of the available credit default prediction models is impaired owing to the fact that most models either fail to identify effectively defaulters, or fail to perform well to new information. Traditions do not generally adequately capture the intricate ways in which various borrower factors, such as as income, employment status, credit score, and loan terms, interrelate [6].

Moreover, there are also challenges like overfittingg, inefficient feature ranking, and lack of sufficient management of class imbalance even when machine learning is used. A complex and all-inclusive framework is needed to increase the total prediction accuracy and reliability, which involves multiple machine learning models, automatizes preprocessing steps, and effectively addresses class imbalance.

Not even modern studies of credited default prediction, such as comparison, scarcely ever discuss the application of ensemble methods, such as stacking, with other techniques of feature selection, balance of classes, and the optimization of hyperversion. There is some limited predictive power in estimating of defaults accurately without compromising the resilience and generalizability of models of such integrated frameworks are unavailable. This study can solve these drawbacks by introducing a stacking ensemble machine learning model predicting loan default. It preoccupiess itself with managing class imbalance, including more features, and modifying hyperparameters.

## 1.3 Research Scope and Approach

The primary objective of this project is to develop a complete machine learning model that forecasts loan defaults in a way that is fully legitimate with the assistance of integrating a number of algorithms into a cohesive strategy. The clear research objectives of this study are:

- •To develop a stacking ensemble model to predict loan defaults: To enhance the precision of predictions, the paper will evaluate the effectiveness of stacking ensemble methods, which involves the use of numerous underlying models, such as Random foresting, XGBoost, ADABoost, SVM and MLP [7].
- To remove the imbalance due to classification: Incorrect resampling result in a reduced study will compare and analyze various techniques of Resampling, including Adaptive Synthetic Sampling (ADASYN), Random Over-Sampling (ROS) and Synthetic Minority Over-Sampling Technique (SMOTE), to reduce the influence of class-imbalancing influence on the model performance [8].
- To maximized the feature selection and preprocessing methods: To discover as many relevant variables as possible and prepare the data to achieve the most successful model performance, the given project will consider a few methods of selecting the features and preprocess it. Some of the common methods of normalization as well as outlier treatments to which we are going to study belong to the Recursive Feature Elimination with Cross-Validation (RFECV) method.
- • To enhance the models with hyperparameter optimization: in this project, the hyperparameters of each of the base models will be varied in order to maximize performance and generalization to new data with methods such as the GridSearchCV.
- Ser: To compare the results of a proposeds model with the current techniques: The application to the proposed model will assess performance using key performance indices, which include accuracy, precision, F1-score, recall, and Area Under the Curve (AUC), and compare them to the baseline methods within similar research areas..

## 1.4 Significance of the Study

The work may have immense influence on both the academic studies and the practical implementation in the financiasl services industry. The key contributions and implications of the research are the following ones:

Enhancing Precision in Loan Default Prophesy: This experiment employs an advanced stacking ensemble model in order to better predict loan defaults. The prediction accuracy is also enhanced by the many base models in stacking structure, especially in the background between the defaultersr and the non-defaulters. This would enhance their credit risk evaluation procedures, reduce the loan loss ratios and enable financial institutions rely on fair judgments [9].

Solving Class Imbalance: The issue of Class imbalance is typical in most financial data. This study contributes to the developments of more efficient measures of the dataset balance and a higher model accuracy on minority classifications through the consideration of various methods of proceding to handle this problem. This will not only ensure the model scales up well to new and unknown data but will also boost the number of defaulters.

Advances in Feature Selection and Data Preprocessing: The discussion on the topic of feature selection methods and methods of data preprocessing, including data normalization and treatmenedt of outliers, will help make the model building process faster. This will ensure that only the most relevant variables are used to predict loan defaults and will also provide essential information on how to properly structure data to machine learning models [10].

Delivering a Future Research Framework: This research-contributing study precursors an extensive new body of knowledged on credit default prediction by including an entire union of methods including ensemble techniques, resampling strategies, feature choice, and hyperparameter tuning. The given methodologies could allow researchers to develop new algorithms or datasets on the basis of the information received in the current study and later refer to it to base the further research.

Practical Implications to Financial institutions: Banks, lending sites, and other related financial institutionss will significantly gain off the findings of the study. Through the proposed methodology, these institutions can enhance their ability to identify high risk borrowers, reduce losses associated with defaults and make superior lending decisions.

This would result in improved risk managements, improved profitability, and increased confidence by the people of financial institutions.

In conclusion, the intended study will offer a more accurated, scalable, and practical solution to the loan default prediction problem. This research will enlarge both theoretical learning and practical implementation of credit risk evaluation systems in financial services industry through complex machine learning algorithms and an ensemble model.

# Chapter 2

## Literature Review

Numerous studies have existed on loan default prediction topic particularly with regard to financial risk managementd. This issue has served as the subject of many different approaches, including ensemble algorithms, machine learning algorithms, and classical statistical models. This literature review offers a thorough overview of the studies carried on loan default prediction, and specifically it focuses on the significant methods, problem that the current models have, and the advantages of applying machine learning methods particularly ensemble learnindg. The main topics of the review are classical models of prediction, machine learning models, methodologies of ensemble learning, and models solving feature selection and class imbalanced.

### 2.1 Traditional Prediction Models

Classically, the rule-based model, logistic regression and other statistical models were employed in epitomizing loan defaults. These strategies generate models based on historical Information that could shine some light on how a borrower is likely to default his loan depending on his financial variables (loan terms, income, employment status, credit score and many others).

Logistic regression is one of the most suitable models in predicting financial risk. It is an overriding, transparent, and widely-used approach to credit rating algorithms. To illustrate, [12] used the logistic regression to predict the likelihood of an individual borrower defaulting a loan based on such parameters as loan amount, credit history, and debt-to-income ratio. The research suggested that logistic regression yielded plausible results but was generally weak due to its inability to address intricate and non-linear relationships in the data.

Decision Trees (DT) is also another outdated technique, whose diagram looks like a flowchart, where each internal node represents a "decision" dependent on a feature and each leaf node the conclusion (a default or non-default). In spite of the fact that the decision trees are easy to learn and understand, as the tree grows deeper, they often become over fit. In at least one study, such as [13], it was demonstrated that decision trees were superior to logistic regression models, but they were still constrained by their lack.

Although simple models, including logistic regression and decision trees can give valuable findings, they are typically insufficient in large datasets that present non-linear relations and complex interactions between features, which is becoming common in financial data

## 2.2 Machine Learning Models

Machine learning has changed the field of loan default prediction by developing more accurate models that can handle complicated, high-dimensional data. Several machine learning algorithms have been applied to credit risk assessment, each having its benefits and shortcomings.

**Random Forest (RF)** is an ensemble method based on decision trees that builds multiple trees during training and outputs the majority prediction. It is less prone to overfitting compared to individual decision trees and has been frequently utilized in credit scoring. In research by [14], random forests surpassed decision trees and logistic regression in terms of accuracy, proving its ability in modeling non-linear connections and managing high-dimensional data. However, random forests can be computationally expensive, especially for huge datasets.

Loan default prediction has also made use of Support Vector Machines (SVM), which look for the hyperplane that best divides data points from distinct classes. For binary classification issues, where the objective is to forecast whether a borrower would default (1) or not (0), SVMs are highly effective. SVMs have proven to be highly accurate in projecting loan defaults, particularly when combined with kernel functions to manage non-linearity, according to [15]. SVMs' performance may deteriorate with very big datasets, though, and they may be sensitive to the kernel and hyperparameter choices.

Extreme Gradient Boosting (XGBoost) is another machine learning algorithms that has a great potential in predicting loan defaults. XGBoost is a gradient booster algorithm, which builds successive sets of decision-trees, with each set of the set covering-up the errors made by the prior set. In other research works, such as [16], XGBoost can be compareds to random forests and the SVM, wherein XGBoost outperforms these algorithms in terms of accuracy, precision, and recall. XGBoost helps the performance on imbalanced datasets as well, given that most loan datasets have significant constituents of non-defaulters, which again demonstrates its usefulness in the context of financial risk prediction.

The Adaptive Boosting (ADABoost) is another algorithm used in boosting that is used to build strong classifier by combinings a large number of weak learners (in most cases decision trees). The action of ADABoost to rectify the errors made by previous classifiers and adjust the weights. In loan default prediction, it has shown that ADABoost, when compared to other models, enhances predictive accuracy through the higher attainment of scrutiny on the more challenging to identify instances. [17] discovered that ADABoost performed well in smaller and simplers datasets, however, failed to tackle overfitting in larger and more involved datasets hence it is better applied in combination with other modeling technique.

Multi-Layer Perceptron (MLPs) is a type of a neural networks which has been utilized in loan default forecasting. MLPs have many layers of neurons connected to each other, and capable of learning such non-linear relationships among the data. Evidence has been presented [18] that MLPs excel common models of machine learning on a large dataset. Nevertheless, MLPs ared also computationally expensive and when not regularized will overfit.

.

# Chapter 3

## Methodology

This chapter presents the research methods and strategy employed in this report to develop a very strong stacking ensemble machine learning model to helps indicate the likelihood of default in a loan. The approach is a detailed, multi-stage procedure, including the steps of data collection and rigorous pre-processing, the steps of class balance mitigations, attentive engineering and selection of the features, and the final production and optimization of an individual and ensemble predictions model. Every step involved a scientific evaluation to concluded the most effective strategies, thus ensuring the reliability and external validity of the proposed solution. The detailed research approach is as illustrated in Figure 1.

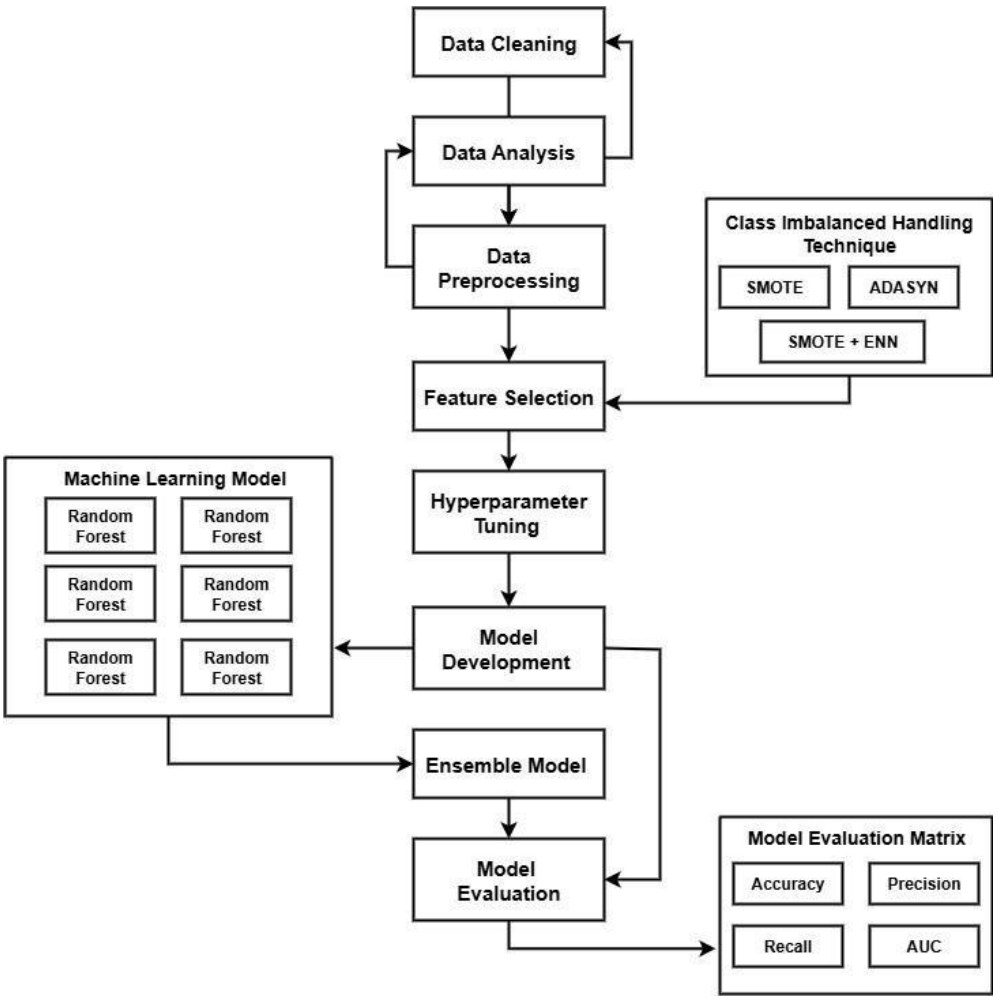



**Figure 1:** Complete Research Methodology Workflow

## 3.1. Data Collection and Characterization

The efficacy of any effective predictions model is contingent upon the quality and representativeness of its foundational data. This paragraph explains the source, collection, and initial characterization of the dataset utilized in this investigation, along with the specific description of the target variable and the sampling approach used.

### 3.1.1. Data Source & Collection

The information to be used in this study was acquireds through LendingClub, one of the leading peer-to-peer (P2P) lenders websites. This site provides detailed information on loans since 2007-2018, which is useful in the assessment of credit risks [19]. The initial dataset involved a huge sample of loan demographics of more than 152 variables with a wide range of borrower users, quality of the loans, and records in their paymenst history [20]. Applicability of LendingClub data is consistent with previous research on credit default prediction, with a comparable basis upon which to evaluate model performance..

### 3.1.2. Definition of the Target Variable

The target variables, y, was also strictly defined to the objectives of such a binary classification task using the loan-status attributed in the LendingClub data set. By restricting the study on certified good and poor loans, they were able to come up with the exact and clear categorization objective. Specifically, in case the status of any loan was Charged off, they were classified under the default category (positive class); when the status was Fully Paid, they were classified unders the non- default category (negative class). This binary categorization is typicallsy used in one Credit default model and that it is possible to directly apply the classification techniques to it..

$$\text{Target(y)} = \begin{cases} 0: & \text{where loan status} = \text{“Fully Paid”} \\ 1: & \text{where loan status} = \text{“Charged off”} \end{cases} \quad (1)$$

### 3.1.3. The stratified sampling method

Given the enormous size of the whole LendingClub dataset, a deliberate sampling technique was vital to balance computational efficiency with data representativeness. A stratified sampling strategy was employed to make sure that the sampled subset still had the same proportion of default and non-default classes as the original dataset [21].

This strategy is particularly crucial in imbalanced datasets to reduce sampling bias towards the dominant class.

Loan default datasets are naturally skewed because there are a lot more non-defaults than defaults. If a basic random sampling technique were applied to such a vast, imbalanced dataset, there would be a substantial possibility that the resultant sampled subset would not accurately reflect the true proportion of the minority class (defaults). This could lead to a sampling dataset that is even more uneven or, in extreme circumstances, altogether lacking examples of the minority class, rendering subsequent modeling efforts worthless for the essential task of identifying defaults. Stratified sampling retains the original class distribution by making sure that each subgroup, such as the default and non-default classes, is represented in the sample in the proper numbers. This preservation is critical for training models that can generalize effectively to both classes, notably the rare but highly important default class. In financial risk modeling, where the penalty associated with misclassifying a defaulter (a false negative) is typically much larger than that of misclassifying a non-defaulter (a false positive), providing proper representation of the minority class from the outset serves as a foundational step. Stratified sample immediately contributes to the model's ability to learn the distinctive characteristics of defaulters, so directly helping the purpose of lowering financial losses. This highlights how vital it is to make a design choice based on the data characteristics of the problem and the overall business goals.

323,025 non-defaults and 80,568 defaults made up the manageable sample size of 403,593 data obtained by stratified sampling [22]. Achieving a compromise between system efficiency and data quality, this precise sample size was created to provide enough data for efficient model training and assessment without depleting cognitive resources.

## 3.2. Exploratory Analysis and Data Pre-processing

Even with sampling, the raw data often has mistaken, inconsistencies, and extraneous characteristics that may seriously impair model performance. To improve the quality of the data and get it ready for further modeling, this step required a painstaking process of data translation, cleaning, and preliminary analysis.

### 3.2.1. Initial Data Cleaning and Handling Missing Values

The data cleaning process was began using a multi-observation cleaning strategy, analogous to the methodology outlined by Ma et al., focused on discovering and rectifying mistakes, inconsistencies, and handling missing values. The dataset was initially examined for duplicate records, none of which were discovered [23].

A noteworthy problem was the occurrence of missing data across 104 characteristics [24]. A realistic technique was designed to solve this:

- **Exclusion of High-Missingness Features:** The analysis thoroughly removed columns with missing data in excess of 50% [25]. This cutoff ensures that features that contain insufficient information won't introduce bias or noise into the model.
- **Removal of Redundant/Non-Informative Categorical Features:** Several categorical features, while having a substantial number of missing values, were also deleted owing to their poor usefulness for the analysis or redundancy.

  Examples include:
  - **emp_title:** This feature contains 142,402 distinct values, making it very granular and possibly noisy for generalized patterns [26]. Its large cardinality would also make one-hot encoding computationally costly and might lead to a sparse, high-dimensional feature space, perhaps creating the curse of dimensionality and overfitting.
  - **title:** With 21,976 distinct values, it was judged similar in aim to the purpose feature, resulting to duplication.
  - **emp_length:** This characteristic exhibited equal poor loan rates across its multiple categories, demonstrating minimal discriminating potential for this particular dataset despite its ubiquitous usage in credit applications. Including such a characteristic would increase noise without much predictive improvement.
- **Imputation for Numerical Features:** To retain crucial information in numerical columns with missing values, an imputation technique was created based on the statistical distribution of each feature, notably its skewness and kurtosis. The Fisher-Pearson coefficient, determined using SciPy, was applied to assess skewness [27].
  - **Median Imputation:** For skewed characteristics, median imputation was employed. The median is resilient to outliers and extreme values, making it an appropriate option for non-normally distributed data, since the mean

would be very susceptible to outliers and skewness, possibly distorting the central tendency.

- **Mode Imputation:** Mode imputationd was also used in multimodal property. The strategy guarantees the most frequent value is selected to preserve the integrity of dominant categories of the data and using thed mean or median may produce an imputed value that is not indicative of the dominant modes.

Such a thorough thinking processs over data drop-outs and feature pruning, well before it puts outliers into perspective or tunes the data, shows that much is understoods of how data quality bias impacts model faithfulness. It encourages fact-fastening and model accuracy beyond targeting the fill-in aspect, a feature of genuine academic scholarly works in data science**.**

### 3.2.2. Multicollinearity Assessment and Feature Removal

**The presence of multicollinearity** when the independent variables of a regressions model have a close association with the rest can cause unstable model coefficients and incorrect conclusions of the resemblance of a feature. To overcome this, Pearson correlation coefficient (r), which is a commonly used form of filtration method of determining linear correlation between variables [28], was used.

$$\text{Correlation } (r) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \quad (2)$$

Characters that amusingly exhibits a Pearson correlation coefficient of more than 90 percent with another character were meticulously identified and eliminated. These stringent requirements ensure that there is not much duplication of information, which causes potential anxieties of inflated standards errors, bad p-values and bad model interpretation. The responsed is known to be consistent with optimal techniques in statistical modeling and machine learning to ensure that the model is robust and will not cause false associations.

### 3.2.3. Exploratory Data Analysis and Outlier Identification

After the cleaning was first completed, the extensive exploratory data analysis (EDA) was conducted to gain a better insight into the properties of the dataset and reveal potential issues that require further processing. All numerical characteristics were generated as

descriptive statistics such as count, mean, median and standard deviation to describe their distribution and find out ab normalities [28].

Data visualization approaches were extensively utilized to visually analyze feature distributions and relationships and help identify those features that provided minimal information or contained errors. 1

As an example, the addr state option, which had been and still represents discrete states, was changed into bigger category of area. This dimension-reduction technique ensures the maximum amount of dimensionality reduction and retained potential regional implications on risk of loan failures. While ***addr_state*** could yield detailed geographical insights, its huge cardinality (many states) might lead to sparse data and overfitting if onehot encoded directly. Aggregating to region decreases dimensionality, perhaps captures larger economic or demographic patterns, and enhances model robustness by lowering sparsity. This indicates an intentional choice to blend specificity with generalizability. The presence of significant outliers and probable data entry errors in critical financial variables like *annual_inc* (annual income) and dti (debt- to-income ratio) was a significant finding of the descriptive analysis. One For instance, the reported highest value of *annual_inc* was GBP 9,522,972, a huge outlier that may indicate an error or a very unrepresentative data point. Such outliers may drastically skew statistical measurements (mean, standard deviation) and affect the learning process of many machines learning models, leading to poor generalization. Similarly, dti indicated a maximum of 999.00%, implying erroneous or exceedingly rare situations. This is especially crucial since an erroneous high-income figure would falsely decrease the DTI, making a high-risk borrower look low-risk, or conversely, an erroneous low income might inflate DTI. This demonstrates a causal chain where mistakes in one feature (annual_inc) spread and distort another derived feature (dti), rendering both unreliable without rectification.

**Table 1**. Statistical summary of annual income and DTI

| Features | Count | Mean | Std | 50% | Max |
|---|---|---|---|---|---|
| **annual_inc** | 403,593 | 76,278.3 | 71,140.2 | 65,000 | 9,522,972 |
| **dti** | 403,593 | 18.26 | 10.38 | 17.62 | 999 |

These extreme results were further validated by visualizations (Figure 2), indicating their propensity to substantially affect model predictions. This section highlights that data pre-processing is not a trivial formality but a key analytical step, including a comprehensive

grasp of the data's domain (finance), its statistical features, and the possible influence of data abnormalities on model fidelity.

The iterative process of cleaning, analyzing, and changing data is crucial to constructing accurate prediction algorithms in real-world settings.

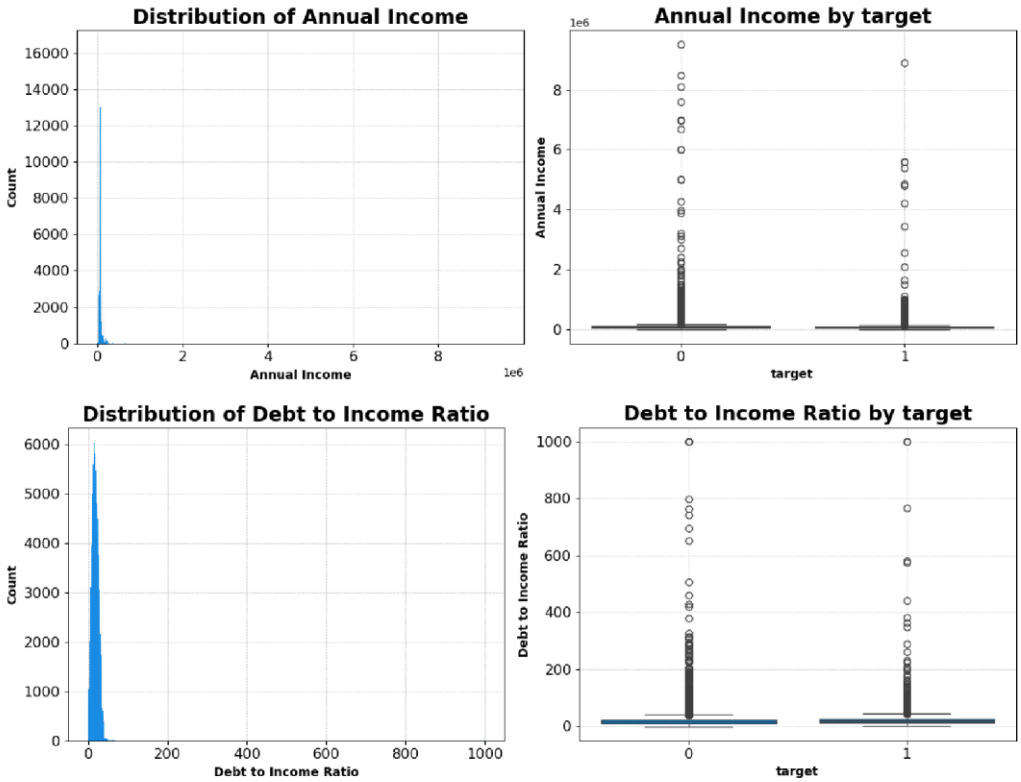



**Figure 2:** Highlights the debt-to-income and annual income ratio (DTI).

### 3.2.4. Outlier Treatment and Data Normalization

The discovered outliers in numerical characteristics were methodically handled to limit their harmful influence on model training. Simultaneously, data normalization was conducted to scale features with divergent ranges, which might otherwise bias certain machine learning algorithms. 1 A systematic assessment approach, leveraging XGBoost as a test model, was created to determine the best mix of outlier treatment and normalization strategies. 1 XGBoost was selected for its efficiency and excellents generalization capabilities, allowing for speedy and accurate evaluation of diverse pre-processing procedures [29].

**Outlier Handling Techniques Evaluated:**

Four separate strategies were evaluated to manage excessive values:

- **Z-score:** This technique uses the calculation $Z=(V-\mu)/\sigma$ to quantify the number of standard deviations a data point (V) deviates from the mean ($\mu$). Severe outliers were defined as values with a Z-score greater than 3.
- **Interquartile Range (IQR):** This robust technique classifies outliers as values lying beyond the boundaries of $Q1-1.5\times IQR$ and $Q3+1.5\times IQR$, where $Q1$ is the 25th percentile and $Q3$ is the 75th percentile.
- **Clip**: This basic procedure limits values below the 1st percentile and above the 99th percentile, thereby truncating extreme values.
- **Winsorize:** Similar to clipping, Winsorization confines extreme values to a defined percentile (e.g., 5th and 95th percentiles), replacing values beyond this range with the values at the stated percentiles [30].

**Data Normalization Techniques Evaluated:**

To ensure features contributed equally to model learning, three standard scaling approaches were tested 1:

- **Standard Scaler:** Transforms data to have a mean of 0 and a standard deviation of 1, following a normal distribution. However, it is vulnerable to outliers.

$$Z = \frac{(V - \mu)}{\sigma} \tag{3}$$

- **Min-Max Scaler:** Rescales data to a defined range, often, using the formula. While less sensitive than the normal scaler, it may still be impacted by extreme values.

$$n = \frac{n - \text{minimum}(n)}{\text{maximum}(n) - \text{minimum}(n)} \tag{4}$$

- **Robust Scaler:** Utilizes the median and the Interquartile Range (IQR) for scaling, making it very resistant to the impact of outliers. The formula:

$$n = \frac{n_i - n_{\text{median}}}{IQR} \tag{5}$$

**Selection of Optimal Pre-processing Strategy:**

The selection of the ideal mix of outlier treatment and normalizing strategies was based on their influence on model performance, principally assessed using recall, precision, and accuracy.

**Table 2.** Finding Outliers and Normalizing Them: Methods of Pre-processing

| Outlier Technique | Normalisation Technique | Accuracy | Recall | Precision | AUC |
|---|---|---|---|---|---|
| z_score | Minmax | 0.7961 | 0.0445 | 0.5444 | 0.6969 |
| z_score | Standard | 0.7963 | 0.0449 | 0.5497 | 0.6971 |
| z_score | Robust | 0.7963 | 0.0449 | 0.5497 | 0.6972 |
| iqr | Minmax | 0.8275 | 0.0035 | 0.5882 | 0.6407 |
| iqr | Standard | 0.8274 | 0.0035 | 0.5263 | 0.6411 |
| iqr | Robust | 0.8274 | 0.0031 | 0.5294 | 0.6410 |
| winsorize | Minmax | 0.8040 | 0.0567 | 0.5544 | 0.7032 |
| winsorize | Standard | 0.8044 | 0.0584 | 0.5625 | 0.7038 |
| winsorize | robust | 0.8045 | 0.0582 | 0.5664 | 0.7039 |
| clip | Minmax | 0.8036 | 0.0544 | 0.5473 | 0.7048 |
| clip | Standard | 0.8040 | 0.0556 | 0.5571 | 0.7049 |
| clip | Robust | 0.8038 | 0.0550 | 0.5516 | 0.7050 |

The results, as shown in Table 2, showed that the best performance metrics were obtained by combining the robust scaler for data normalization with the winsorize strategy for outlier treatment. This was especially true for the XGBoost test model in terms of recall (0.0582) and accuracy (0.5664). The choices of winsorize for outlier treatment and robust scaler for normalization indicates a clever approach. Instead of overtly deleting outliers, which may cause to data loss, winsorization limits them at a defined percentile.

This keeps the data points while decreasing their extreme effect, which is critical when every data point, even an outlier, could include some useful information about the underlying process, or when data scarcity makes removal undesirable. The Robust Scaler was especially selected owing to its inherent robustness to outliers, since it utilizes the median and IQR, which are less impacted by extreme values than the mean and standard deviation. This characteristic comes in particularly handy with financials data that are often affected by huge outliers. Winsorize norms the biggest possible values followed by robust scaler ensure that the subsequenst (capped) extreme values and the rest of the data are scaled appropriately without them being distorted with the initial extremes.

It is a two-step methods providing a very robust pre-processing pipeline to real-world financial data that is noisy. This methodology choice indicates good understanding of the characteristics of data in financial data sets toward robust and information protection, which are most relevant when data integrity has a direct impact on financial outcomes. The use of XGBoost as a testbed for choosing these strategies further indicates an efficient

and data-driven approach to optimization. The modified distributions ***of annual_inc*** and dti following outlier treatment are graphically displayed in Figure 3.

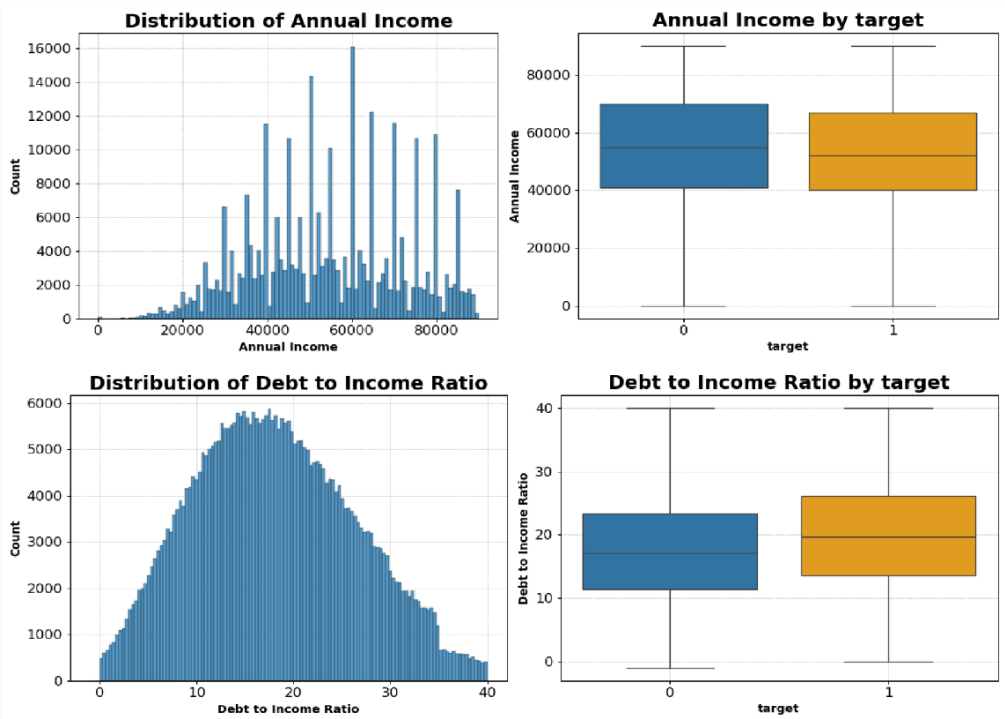



**Figure 3:** Visual inspection of features after addressing outliers

## 3.3. Addressing Class Imbalance

In domains like credit risk prediction, where the incidence of the minority class (loan defaults) is much lower than that of the majority class (non-defaults), class imbalance is a persistent and significant challenge in predictive modeling. This intrinsic imbalance may result in biased models that forecast the majority class with high overall accuracy while performing appallingly poorly on the minority class, which is the class that is primarily of interest for risk reduction.

### 3.3.1. Overview of Class Imbalance in Loan Default Data

As emphasized in the first pre-processing phases (Section 3.2.4), the model's accuracy, although seeming high, did not appropriately represent its performance on the minority class, as demonstrated by much lower recall values [31]. This disparity underlines the deceiving nature of accuracy in unbalanced datasets. The LendingClub dataset included in this research clearly highlights this problem, with non-defaults totaling around 80% of the data, and defaults just 20%. This substantial imbalance, graphically shown in Figure

4, demands specialized measures to guarantee models are not skewed towards the over-represented subgroup.

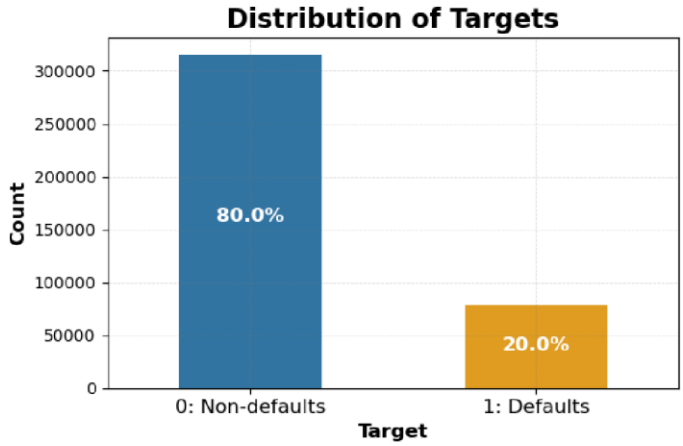



**Figure 4:** Analysis of Target Distribution

### 3.3.2. Resampling Techniques Evaluated

To offset the unfavorable impacts of class imbalance, a full array of resampling strategies was thoroughly examined. No single strategy is generally optimum; hence a comparative study was needed to discover the best suited method for this unique dataset. The methodologies investigated, as indicated in Figure 5, fall into three broad categories: over-sampling, under-sampling, and hybrid approaches.

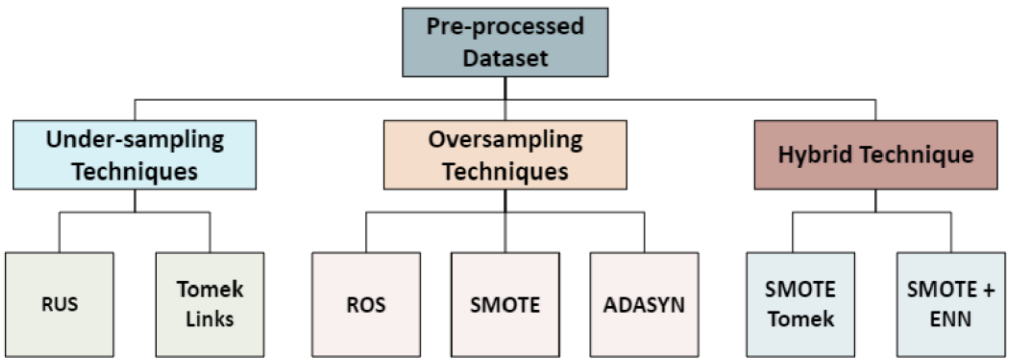



**Figure 5:** Approaches for Handling Class Imbalance The examined approaches include:

1. **Random Over-Sampling (ROS):** Using this method, instances from the minority class are replicated at random until the dataset is balanced. While easy, ROS poses

the danger of overfitting, since it makes exact replicas, possibly forcing the model to learn particular cases rather than generalizable patterns.

$$\text{New } S_{minority} = S_{minority} \cup \{S_{minority} \text{ duplicated until } |S_{minority}| = N_{majority}\} \quad (6)$$

2. **Random Under-Sampling (RUS):** Conversely, RUS balances the dataset by randomly eliminating instances from the dominant class [32]. Although it mitigates overfitting difficulties associated with ROS, RUS may lead to a large loss of potentially essential information from the majority class, which can impair the model's capacity to learn complete patterns.

$$\text{New } S_{majority} = S_{minority} \cup \{S_{majority} \text{ duplicated until } |S_{majority}| = N_{minority}\} \quad (7)$$

3. **Synthetic Minority Over-sampling Technique (SMOTE):** SMOTE creates synthetic minority class samples by interpolating between existing minority class instances and their k-nearest neighbors [33]. This strategy provides fresh, unique data points, so enriching the minority class without merely replicating existing ones, which helps in avoiding overfitting.

$$x_{new} = x_i + \nearrow \times (x_{nn} - x_i) \quad (8)$$

**Adaptive Synthetic Sampling (ADASYN):** ADASYN is a variation of SMOTE that synthetically oversamples additional samples of cases of a minority class that are more difficult to classify (i.e. have more majority class neighbors). 1 This more targeted generation methods aims to push the decision boundary in favor of the minority class, minimizing the classification error.

$$G_i = d_i \times G \quad (9)$$

4. **Tomek Links:** This is an under-sampling approachs used for data cleansing, especially successful around decision boundaries. A Tomek Link consists of a pair of instances from distinct classes that are closest neighbors to each other.

Removing the majority class instance from such a pair helps to clarify the decision boundary and reduce noise.

$$\text{Remove } (x_1, x_2) \begin{cases} \text{if } x_1 \text{ and } x_2 \text{ are nearest neighbours} \\ + \\ \text{if } x_1 \text{ and } x_2 \text{ belong to different classes} \end{cases} \quad (10)$$

5. **SMOTE-Tomek:** This hybrid technique combines the strengths of over-sampling and under-sampling [34]. First, SMOTE is utilized to produce synthetic minority samples. Subsequently, Tomek Links are discovered and eliminated from the combined dataset, thus clearing the decision boundary and enhancing the quality of the synthetic data.
6. **SMOTE with Edited Nearest Neighbours (SMOTE+ENN):** Another effective hybrid strategy is SMOTE+ENN. Using SMOTE [35], it first oversamples the minority class. After then, both synthetic and original samples are subjected to the Edited Nearest Neighbor (ENN) criteria. Among its k-nearest neighbors, ENN removes instances (from both majority and minority classes) whose class label differs from the majority class. By effectively removing noisy or unclear data points, this "cleaning" procedure lowers the decision boundary and improves the quality of the balanced dataset.

ENN cleaning:

$$|\text{If } x_i \text{ is misclassified by its k nearest neighbours, remove } x_i|$$

This technique can be represented as follows:

$$S_{balanced} = ENN(SMOTE(S_{minority}, S_{majority})) \quad (11)$$

### 3.3.3. Selection of Optimal Class Balancing Method

The different resampling strategies were carefully tested using XGBoost, consistent with its function as the test model for pre-processing stages, and rated based on accuracy, precision, recall, and AUC. The findings of this comparative study are reported in Table 5.

**Table 5.** Implementation of Sampling Methods

| Method | Accuracy | Precision | Recall | AUC |
|---|---|---|---|---|
| None | 0.8047 | 0.5362 | 0.1101 | 0.7171 |
| ROS | 0.6874 | 0.6807 | 0.7062 | 0.7559 |
| SMOTE | 0.8766 | 0.9684 | 0.7787 | 0.9284 |
| ADASYN | 0.8745 | **0.9686** | 0.7690 | 0.9266 |
| RUS | 0.6500 | 0.6465 | 0.6683 | 0.7079 |
| Tomek-Links | 0.7947 | 0.5368 | 0.1377 | 0.7197 |
| SMOTE-Tomek | 0.8762 | 0.9679 | 0.7779 | 0.9295 |
| **SMOTE + ENN** | **0.9049** | 0.9461 | **0.9202** | **0.9654** |

The empirical results clearly showed that, on all metrics, balanced datasets performed much better than the initial unbalanced sample. With an accuracy of 90.49%, precision of 94.61%, recall of 92.02%, and AUC of 96.54%, SMOTE+ENN had the best overall performance among the studied techniques.

The enhanced performance of SMOTE+ENN may be due to its dual capacity. While simple oversampling strategies like ROS or SMOTE boost minority class representation, they might add noise or produce samples too near to the decision boundary, possibly leading to overfitting or sub-optimal bounds. Similarly, simple under sampling approaches like RUS or Tomek Links lower the majority class, which might lead to a loss of crucial information, particularly if the majority class is varied. SMOTE+ENN solves these issues by not only successfully balancing the dataset via synthetic minority sample creation (by SMOTE) but also concurrently cleaning the data by eliminating noisy or confusing cases around the decision border (with ENN). This combination strategy guarantees that the model learns from a cleaner, more representative dataset that has better class separation, leading to much enhanced generalization. The high recall of 92.02% attained by SMOTE+ENN is especially remarkable, as it shows the model's excellent capacity to properly identify the minority class (defaulters), an important aim in credit risk prediction where the penalty of false negatives (missing a defaulter) is substantial. Consequently, SMOTE+ENN was chosen as the ideal strategy for balancing the LendingClub dataset for all further model development [35]. After balancing, the dataset was divided into an 80:20 ratio for training and testing, respectively, a standard method in machine learning validation. This illustrates a mature approach to tackling data difficulties, going beyond simple solutions to a more sophisticated awareness that data quality (cleanliness, boundary clarity) is as crucial as data quantity (balance) for successful machine learning, especially in high-stakes areas like finance.

## 3.4. Feature Engineering and Selection

Enhancing model performance and interpretability, especially in complex datasets, requires careful consideration of feature engineering and selection. In order to maximize predictive value while reducing computing overhead, this step included converting existing characteristics and methodically determining the most relevant subset of information.

### 3.4.1. Transformation and Feature Engineering

In order to capture non-linear correlations and aggregate data, some aspects were further designed or binned after the initial cleaning and one-hot encoding of categorical variables (as shown in Table 2). Annual_inc and revol_bal, for example, were categorized as "Very Low," "Low," "Medium," "High," and "Very High." This binning may change continuous data into ordinal ones, making them more resilient to outliers and perhaps showing non-linear patterns that linear models would overlook.

### 3.4.2. Cross-Validation and Recursive Feature Elimination (RFECV)

The Recursive Feature Elimination with Cross-Validation (RFECV) wrapper approach was used to decrease dimensionality and find the most influential features. In order to find the ideal subset of features, RFECV is an iterative process that methodically eliminates features, constructs a model, and assesses its performance.

Reducing financial losses from loan defaults is the main goal of financial lending. This implies that the institution will suffer large financial losses if a defaulting borrower is mistakenly classified as a non-defaulter (a false negative). A false positive, on the other hand, could result in a missed revenue opportunity, which is usually less expensive than an actual default. Therefore, recall was specifically selected as the scoring criteria for RFECV due to the critical need of accurately recognizing loan defaults to minimize financial losses. Recall is a direct indicator of the model's capacity to identify real positive instances, or defaulters; a high recall means that fewer real defaulters are overlooked [36]. The model is directed to prioritize characteristics that are best at detecting all real defaulters by optimizing for recall during feature selection, even if this results in a small increase in false positives. A hallmark of applied machine learning in high-stakes domains is this deliberate selection of evaluation metrics, which shows that successful model building involves more than just optimizing a single metric; it also entails matching model performance to actual business goals and the asymmetric costs associated with various error types.

To enable a more detailed search for the ideal feature subset, the ***step*** parameter for RFECV was set to 1, which meant that one feature was eliminated in each iteration. In order to balance model complexity with predictive performance, the procedure sought to determine the smallest number of characteristics that produced the greatest recall score. Figure 6 displays the RFECV process results, showing how the number of features and

cross-validation recall score relate to one another. With an ideal recall score of 92.16%, 48 characteristics were found to be the ideal number.

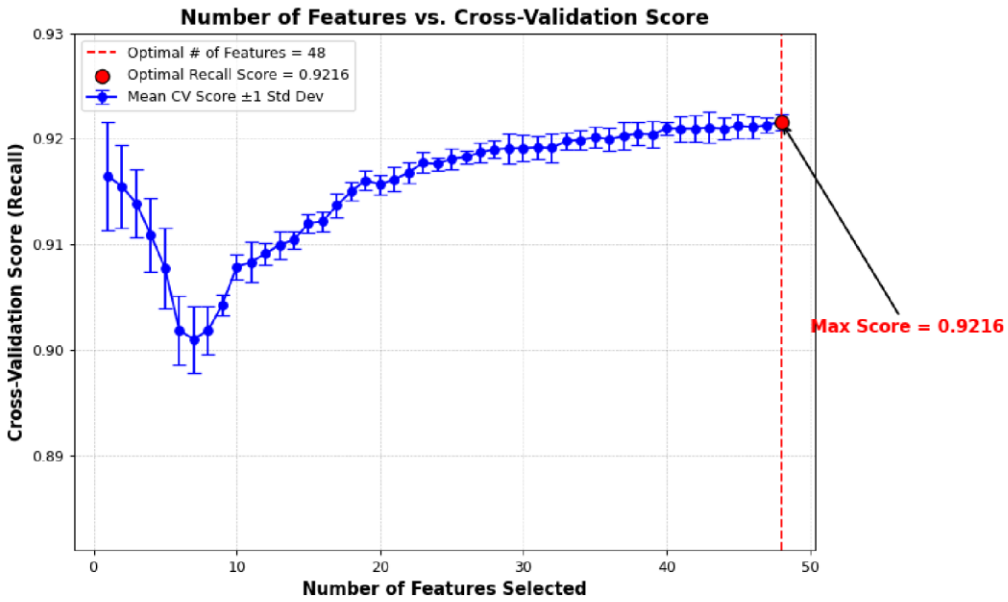



**Figure 6:** Selected Relevant Features

### 3.4.3. Feature Importance Analysis

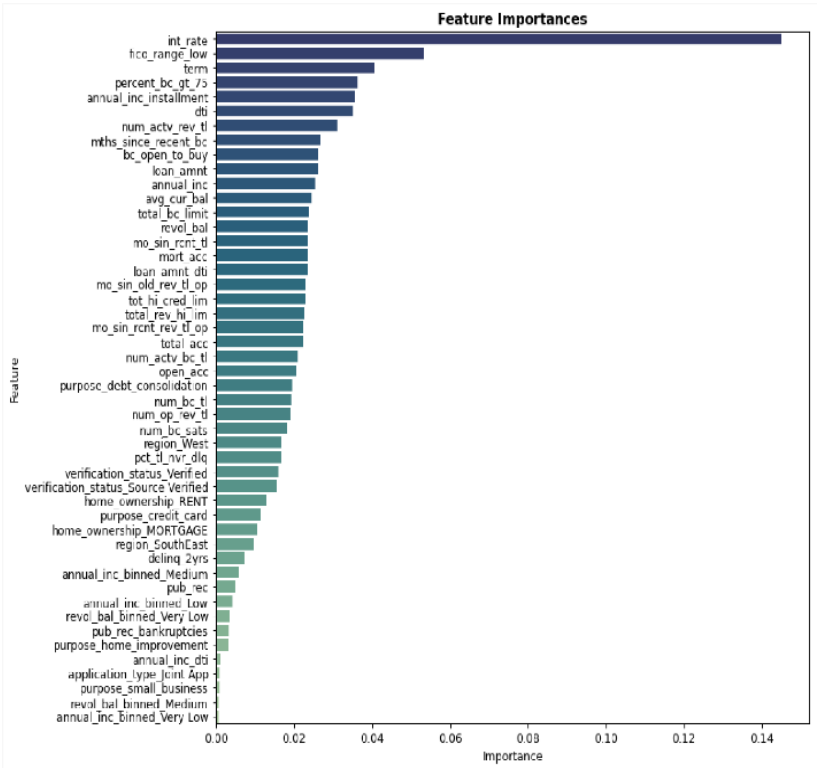



**Figure 7:** Importance of Features

Complementary to RFECV, a feature significance analysis was undertaken to assess the individual contribution of each chosen feature to the prediction model. This research gives useful insights into the fundamental determinants of loan default risk. As indicated in Figure 7, the most important characteristics discovered were the int_rate (interest rate), fico_range_low (borrower's lower FICO boundary range/credit score),[37] and term (loan duration). These results correspond with financial sense, since interest rates and credit ratings are straightforward indications of a borrower's risk profile, while loan length determines the duration of exposure to risk.

## 3.5. Model Development and Improvement

After comprehensive pre-processing, balancing, and feature selection of the data, the next phase involved the construction and optimization of many machine learning models, culminating in the development of a robust stacking ensemble.

### 3.5.1. Use of Individual Predictive Models

To predict credit loan defaults, a wide range of distinct machine learning and deep learning models were used as foundation learners. A variety of algorithmic paradigms were used in the selection process in order to capture various facets of the underlying patterns in the data. Below is a brief description of the theoretical foundations and working mechanics of each model:

A. **Decision Tree (DT):** A non-parametric supervised learning technique that builds a model of decisions and their potential outcomes in the shape of a tree. In order to produce homogenous subsets, it recursively divides the data according to decision rules that are derived from features. The Gini impurity criteria, which calculates the likelihood of incorrectly classifying a randomly selected element if it were randomly labeled based on the node's label distribution, was applied in this work to separate nodes [38].

$$\text{Gini} = 1 - \sum_{i=1}^{c} (p_i)^2 \tag{12}$$

B. **Random Forest (RF):** An ensemble learning technique that constructs several decision trees during training and produces a class that is the mean prediction

(regression) or the mode of the classes (classification) of the individual trees. Compared to a single decision tree, Random Forest increases generalization and decreases overfitting by combining predictions from several trees. Additionally, it splits using the Gini index.

C. **Support Vector Machine (SVM):** An effective supervised learning model for tasks including regression and classification. SVMs seek to identify the best hyperplane in a high-dimensional space that maximally divides data points of various classes. Kernel functions, such as linear, polynomial, and Radial Basis Function (RBF), are used to convert non-linearly separable data into a higher dimension where linear separation is feasible.

$$h(x_i) = \text{sign}(w \cdot x_i + b) \tag{13}$$

D. **Extreme Gradient Boosting (XGBoost):**

An extremely effective and adaptable open-source implementation of the gradient boosting framework is Extreme Gradient Boosting (XGBoost). In a sequential fashion, XGBoost iteratively constructs an ensemble of weak prediction models, usually decision trees, where each new model fixes the mistakes of the ones before it. To improve generalization and avoid overfitting, it uses regularization terms (L1 and L2).

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(X_i)\right) + \Omega(f_t) \tag{14}$$

E. **Adaptive Boosting (ADABoost):**

Another well-liked boosting technique that combines several "weak" learners to produce a "strong" learner is called Adaptive Boosting (ADABoost). Weak classifiers are iteratively trained by ADABoost, which gives misclassified cases bigger weights in later rounds. The weighted aggregate of the forecasts made by the weak learners makes up the final forecast.

$$H(x) = sign\left[\sum_{t=1}^{T} \alpha_t h_t(x)\right] \tag{15}$$

F. **Multi-Layered Perceptron (MLP):**

A type of feedforward Artificial Neural Network (ANN) that consists of an input layer, one or more hidden layers, and an output layer is called a multi-layered perceptron (MLP). Every neuron in a layer has connections to every other neuron in the layer below it, and these connections have weights attached to them. The

MLP may learn intricate patterns thanks to the introduction of non-linearity brought about by non-linear activation functions, such as the sigmoid function for the binary classification output layer and the Rectified Linear Unit (ReLU) for hidden layers. In order to modify weights, the model is trained using backpropagation by minimizing a loss function, usually binary cross-entropy for binary classification.

$$L = -\frac{1}{n}\sum_{i=1}^{n}[y_i log\ (\hat{y}_i) + (1 - y_i)\ log(1 - y_i)\ ] \qquad (16)$$

### 3.5.2. Hyperparameter Tuning using GridSearchCV

A methodical hyperparameter tuning procedure was carried out using GridSearchCV in order to maximize the performance of each distinct prediction model and improve their ability to generalize to unknown data. In order to find the ideal configuration that produces the highest performance based on a given metric, GridSearchCV thoroughly searches through a predetermined set of hyperparameter values for each model, assessing each combination through cross-validation.

In order to balance the trade-off between model complexity and the possibility of overfitting, the ranges of hyperparameters for tuning were carefully chosen. To influence bias-variance trade-offs, for example, ***max_depth*** and ***n_estimators*** were selected for tree-based models (Decision Tree, Random Forest, XGBoost, and ADABoost) in order to manage ensemble size and tree complexity, respectively. parameters such as The variables ***min_samples_split*** and ***min_samples_leaf*** Decision trees were chosen in order to prevent splits with insufficient samples and prevent overfitting. The C parameter (strength of regularization) and the types of kernels were searched to identify the best decision boundary and the feature space transformation of SVMs. MLP parameters that were selected to ensure the model could fit both complex patterns and not get too complicated include the C parameter to trade-off between greater generalization (better fit through

more training data) and smaller learning rates that would help the model understand its preferences without using adversely complex cost functionalities.

The discovery of non-linear relations among data is made possible by the non-linear nature introduced by activations functions; hidden_layer_sizes determines the capacity of the model to develop complex functions. It exhibits in-depth understanding of the fundamental operation of algorithms and their interaction with data characteristics, rather than superficial application of algorithms.
In complex prediction tasks such as loan default to have state-of-the-art performance, this laborious tuning is necessary. Specifics of the hyperparameter grids which were explored are given in table A2 (appv02).

### 3.5.3. Ensemble Learning Strategies: Voting and Stacking

An important principle of this work was ensembled learning involved combining multiple individual learners to generate more effective predictive behaviour than by using any particular model. The basic principles is that the inefficiency of each model can be overcome by the combined effectiveness of a number of different models making predictions that would be more accurate and dependabled. The two primary methods of ensemble that were explored were soft voting and stacking.

Figure 8 shows the general comprehension of ensemble approache

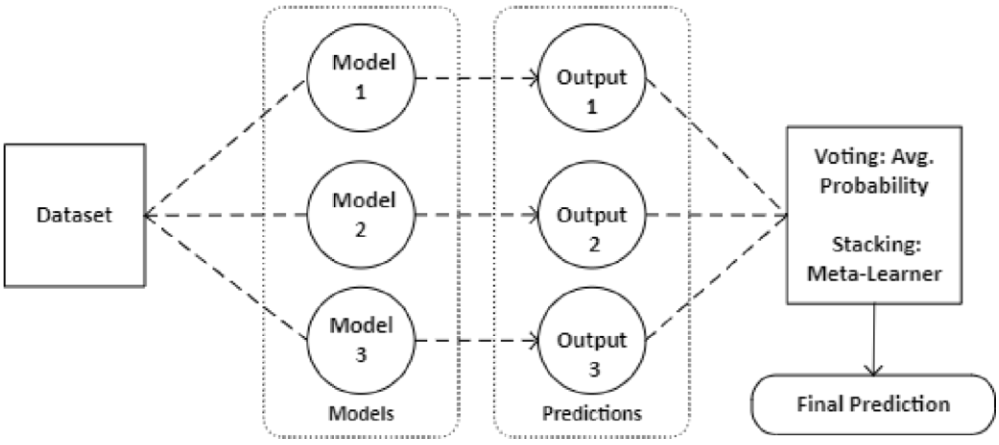



**Figure 8:** Ensemble Modeling Techniques

A. **Soft Voting:** This technique averages the projected odds for each class across predictions from several base models. The final prediction is then made for the

class with the highest average probability among all base learners. There were two types of soft voting used:

- **Method A (Voting A):** Combining the predictions made by the six different models (MLP, XGBoost, Decision Tree, Random Forest, SVM, and ADABoost).
- **Method B (Voting B):** Combining forecasts from the top three individual models, as determined by evaluating their performances separately (Section 4.1.1) [40].

B. **Stacking:** A meta-model, also known as a meta-learner, is used in stacking, a more sophisticated ensemble technique, to determine the best way to aggregate the predictions of several base models. The meta-model then generates the final prediction using the predictions (or probabilities) from the underlying models as input characteristics. This enables the meta-model to determine the advantages and disadvantages of every base learner and to either combine their outputs in a non-linear fashion or apply suitable weights [41]. The meta-model can learn to balance these strengths, for example, if one base model is very good at detecting high-risk defaults but generates a lot of false positives, while another is more cautious and has fewer false positives but misses a lot of actual defaults. Because of its ease of use and interpretability when mixing probability, Logistic Regression was selected as the meta-model for this investigation. Two different stacking configurations were used:

- **Method A (Stacking A):** The predictions from all six separate models were combined using Logistic Regression as the meta-model.
- **Method B (Stacking B):** The predictions from the three top-performing individual models were combined using Logistic Regression as the meta-model.

In order to empirically ascertain whether a more focused mix of high-performing models or a wider variety of models produces better results in the context of loan default prediction, both the "all models" and "best performing models" options were tested. A simple average may not be adequate in extremely complicated and unbalanced areas like financial risk, as evidenced by the emphasis on stacking, especially "Stacking A" (combining all models). From the "opinions" of the underlying models, a meta-learner can extract deeper, higher-order patterns, resulting in a more complex and ultimately more

accurate final judgment. This pushes predictive modeling's limits to intelligent model collaboration rather than individual model optimization.

## 3.6. Model Evaluation Metrics

The full evaluation of model performance is critical, especially in the setting of imbalanced datasets like loan default prediction, where traditional measures might be misleading. To give a holistic and reliable assessment of the models developed, a set of widely recognized evaluation indicators was applied. These measures were also essential in the selection of suitable pre-processing strategies (Section 3.2.4) and class balancing approaches (Section 3.3.3).

The major evaluation metrics employed are:

1. **Accuracy:** Measures the overall proportion of correctly identified occurrences (including true positives and true negatives) out of the total number of predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FT + FN} \tag{17}$$

   **Relevance for Imbalanced Datasets:** While often utilized, accuracy might be deceiving in imbalanced datasets. A model might attain high accuracy by merely predicting the majority class for all cases, disguising weak performance on the minority class. This tendency was specifically noted in the study's initial assessments, where high accuracy did not correspond with great memory for the minority class.

2. **Precision:** Quantifies the proportion of real positive predictions among all instances projected as positive. In the context of loan default, precision reflects the dependability of the model's positive predictions (i.e., when the model predicts a default, how likely is it to be an actual default).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{18}$$

   **Relevance for Imbalanced Datasets:** Precision is crucial as it directly tackles the cost of false positives. High precision means fewer healthy borrowers are falsely classified as defaulters, which is vital for maintaining consumer trust and preventing lost business prospects.

3. **Recall (Sensitivity or True Positive Rate - TPR):** Measures the proportion of genuine positive instances that are accurately detected by the model. In loan

default prediction, recall signifies the model's capacity to identify all real defaulters.

$$\text{Recall(Sensitivity)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{19}$$

**Relevance for Imbalanced Datasets:** Recall is likely the most significant statistic for imbalanced datasets in financial risk, especially when the cost of missing a positive instance (a defaulter, i.e., a false negative) is exceptionally high. In financial lending, the financial cost of a false negative (an actual default) is often significantly more than the cost of a false positive (a wasted opportunity to lend). A strong recall directly corresponds to avoiding financial losses by guaranteeing that most true defaulters are recognized and appropriate risk mitigation methods may be employed [42]. The study deliberately prioritized recollection due to this business requirement. This strategic choice of evaluation metrics is a hallmark of applied machine learning in high-stakes domains, demonstrating that effective model building is not just about maximizing a single metric but about aligning model performance with real-world business objectives and the asymmetric costs associated with different types of errors.

4. **Area Under the Receiver Operating Characteristic Curve (AUC-ROC):** The AUC-ROC quantifies the model's ability to differentiate between positive and negative classes across all potential classification levels. The ROC curve compares the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold levels.

   **Relevance for Imbalanced Datasets:** Because it provides a single, aggregate performance metric that is less sensitive to class distribution than accuracy, AUC is especially significant in datasets that are unbalanced. Regardless of the predetermined classification threshold, a higher AUC indicates a greater ability of the model to distinguish between defaulters and non-defaulters [43].

# Chapter 4

## Results and Discussion

This section contains the empirical data gained from the systematic methodology indicated in Section 3. The findings involve the effects of correcting class imbalance, the feature selection process, and the performance evaluation of both solo and ensemble machine learning models. These data are later compared with established baseline investigations to validate the efficacy of the suggested technique.

### 4.1 Model Performance Evaluation

This chapter shows the performance evaluation of the individual base learners and the proposed ensemble models, applying the comprehensive set of measures outlined in Section 3.6.

#### 4.1.1 Individual Model Performance

The performance of the six separate machine learning and deep learning models, serving as base learners, is reported in Table 6.

**Table 6**. Performance of Individual Model

| Model | Accuracy | Precision | Recall | AUC |
|---|---|---|---|---|
| Random Forest * | 0.8987 | 0.8996 | 0.9656 | 0.9589 |
| Decision Tree | 0.7778 | 0.7743 | **0.9713** | 0.7256 |
| SVM | 0.7318 | 0.9476 | 0.6601 | 0.8824 |
| XGBoost * | **0.9156** | **0.9478** | 0.9330 | **0.9726** |
| ADABoost | 0.8458 | 0.8548 | 0.9439 | 0.9305 |
| MLP * | 0.8775 | 0.9008 | 0.9305 | 0.9229 |

* Indicates models' part of the ensemble with 3 base-learners.

Table 6 shows that ensemble-based models, such Random Forest and XGBoost, often performed better than more straightforward models, including Decision Tree and Support Vector Machine (SVM). With an accuracy of 91.56%, precision of 94.78%, recall of 93.30%, and AUC of 97.26%, XGBoost showed the best individual performance. The strong gradient boosting architecture of XGBoost, which repeatedly fixes mistakes and incorporates regularization to avoid overfitting, may be the cause of this improved performance. It is particularly useful for complex financial datasets. Random Forest likewise demonstrated great performance with a recall of 96.56% and an AUC of 95.89%.

With a recall of 93.05%, the Multi-Layered Perceptron (MLP) also performed well. Despite having a high accuracy value, the SVM performed the worst overall, demonstrating its limitations in identifying the subtle non-linear correlations seen in credit default data [44].

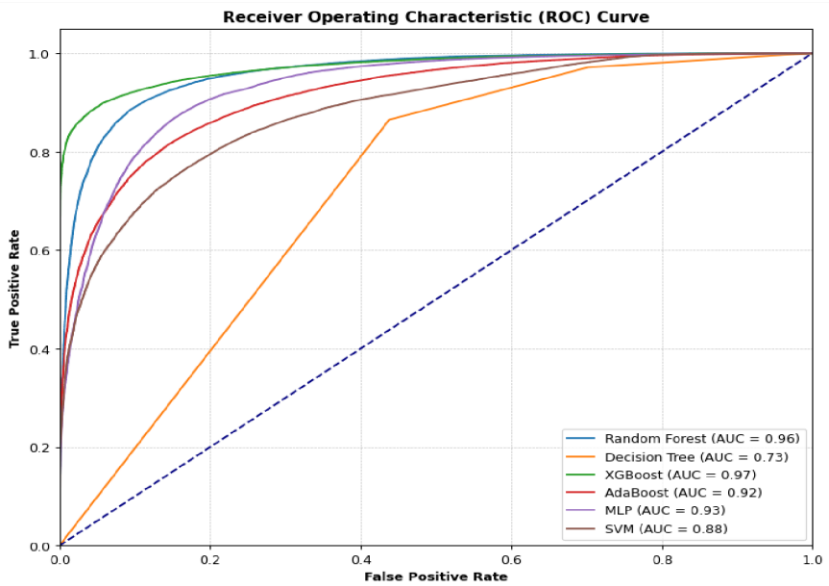



**Figure 9:** ROC Analysis of Individual Predictive Models

Figure 9, shows the separate models' Receiver Operating Characteristic (ROC) curves, which visually validate their discriminative power across a range of classification criteria. XGBoost had the highest area under the curve (AUC), demonstrating its superior overall discriminative capability, but other models showed a respectable ability to distinguish between default and non-default classes. A comparison of these unique model performances across all metrics is shown in Figure 10.

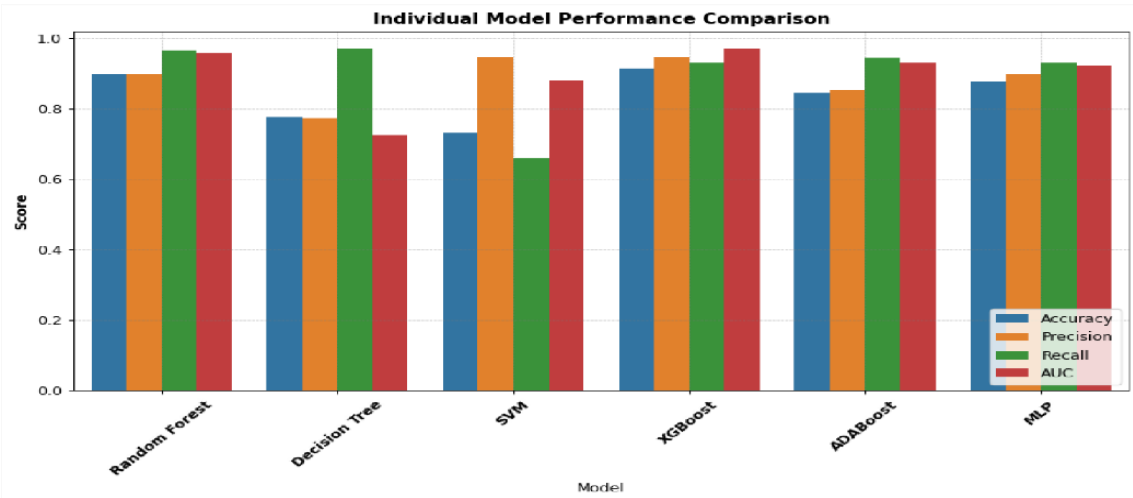



**Figure 10:** Performance Evaluation of Individual

### 4.1.2. Performance of Ensemble Models

To further improve predicted accuracy and resilience, ensemble learning techniques—soft voting and stacking—were used to build on the individual model results. Table 7 displays the outcomes of these ensemble approaches.

**Table 7**. Performance of Ensemble Model

| Model | Accuracy | Precision | Recall | AUC |
|---|---|---|---|---|
| Voting A | 0.9109 | 0.9099 | **0.9710** | 0.9703 |
| Voting B | 0.9166 | 0.9314 | 0.9532 | 0.9687 |
| Stacking A | **0.9369** | **0.9559** | 0.9555 | **0.9781** |
| Stacking B | 0.9188 | 0.9409 | 0.9454 | 0.9708 |

Table 7's findings unequivocally show that, when compared to individual models, integrating the predictions from several models greatly enhanced overall performance. Among the group methods. Stacking The most striking results were obtained with an approach that combined predictions from all six separate base learners using Logistic Regression as a meta-model. The results for **Stacking A** were 93.69% accuracy, 95.59% precision, 95.55% recall, and 97.81% AUC.

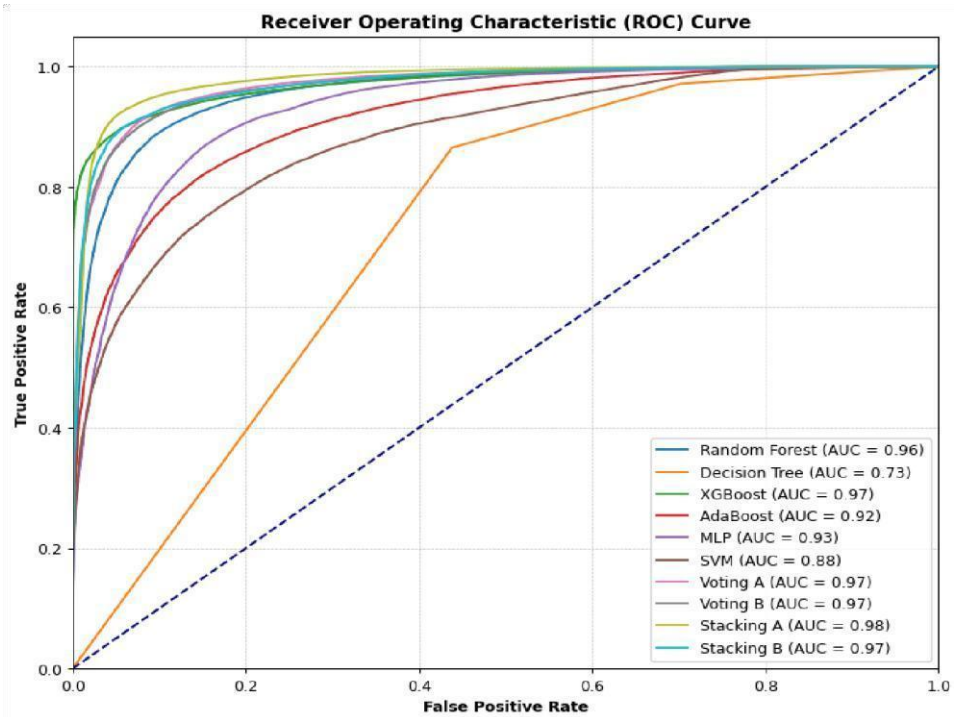



**Figure 11:** ROC Analysis of all Models

This significant improvement highlights the effectiveness of stacking, in which the meta-model learns to mix and weigh the strengths of various base learners in the best possible way to provide a more accurate and nuanced final prediction. The ensemble models' better discriminative ability is further demonstrated by the ROC curves in Figure 11, where Stacking A exhibits the greatest AUC, indicating its remarkable ability to distinguish between defaulters and non-defaulters. A thorough comparison of all models is shown in Figure 10, emphasizing Stacking A's superior performance [45].

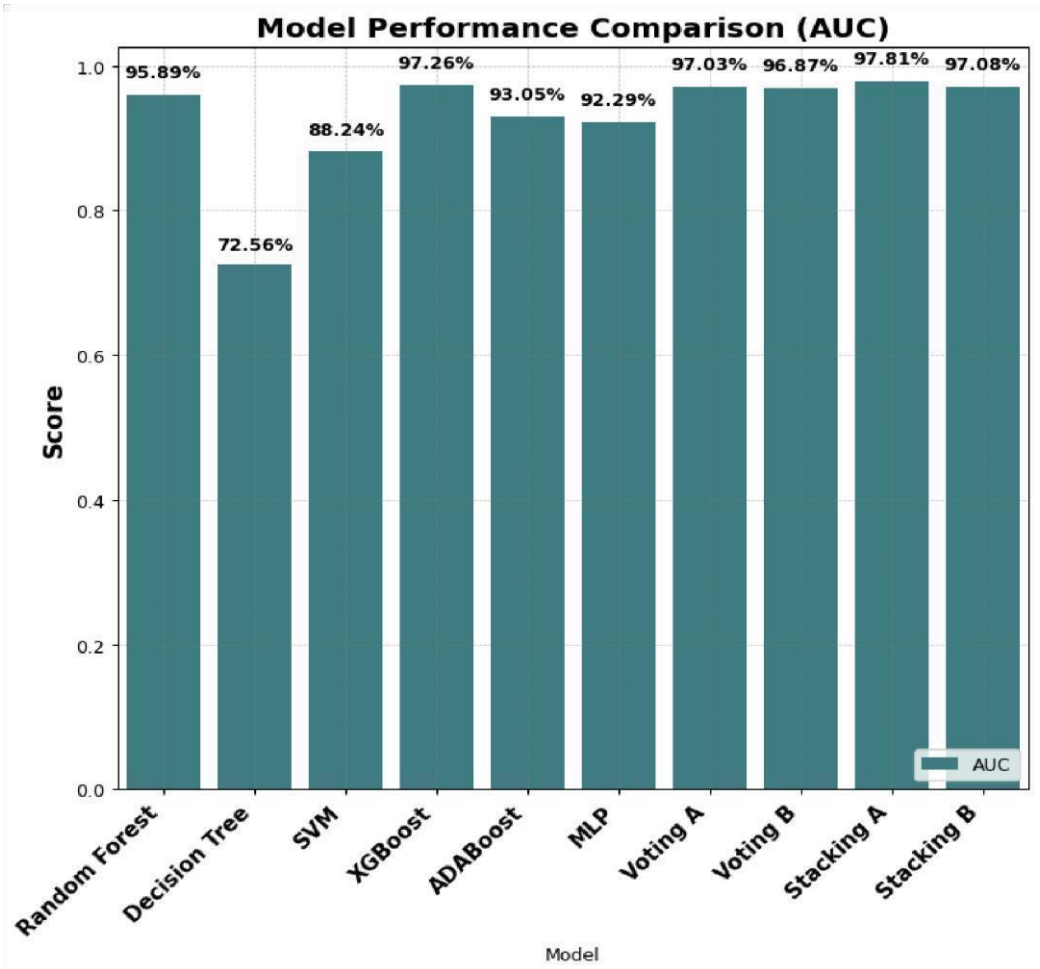



**Figure 10.** Overall Model Performance Comparison

Because they demonstrate the model's strong reliability in making positive predictions and its capacity to accurately identify the vast majority of actual defaulters—a crucial goal in financial risk management—Stacking A's high precision and recall values (roughly 96%) are especially noteworthy. This result supports the theory that intelligent cooperation between various models, enabled by a meta-learner, can result in cutting-edge performance in challenging classification tasks such as loan default prediction.

## 4.2. Comparative Analysis with Baseline Studies

The results of the suggested stacking ensemble model were thoroughly compared to a number of pertinent baseline studies that used comparable LendingClub datasets for credit default prediction in order to put its performance into context. This comparative analysis is summarized in Table 8.

**Table 8**. Comparison with Baseline Model

| Reference | Class Imbalance Method | Models | Ensemble Technique | Data Split | Performance Metrics | Best Model | Best Model Score |
|---|---|---|---|---|---|---|---|
| **This Study** | **ROS, RUS, SMOTE, ADASYN, Tomek Links, SMOTE-Tomek, SMOTE + ENN** | **Random Forest, Decision Tree, SVM, XGBoost, ADABoost, MLP (Three Hidden Layer)** | **Voting, Stacking** | **80:20** | **Accuracy, Precision, Recall, AUC** | **Stacking Based Model** | **94%, 96%, 96%, 98%** |
| [1] | None | Random Forest, Decision Tree | None | 70:30 | Accuracy | Random Forest | 80% |
| [27] | None | LightGBM, XGBoost | None | 91:9 | Accuracy, Error Rate | LightGBM | 80%, 20% |
| [29] | Cluster-based under-sampling | Logistic Regression, SVM, XGBoost, Group Method of Data Handling | None | 80:20 | Accuracy, AUC | XGBoost | 90%, 94% |
| [33] | SMOTE | Logistic Regression, Decision Tree, SVM, ADABoost, MLP (one-hidden layer), MLP (three hidden layer) | None | 80:20 | Accuracy | MLP (three hidden layer) | 93% |

The suggested stacking ensemble model (Stacking A) performed noticeably better than the baseline models on every evaluation criterion, as shown in Table 8. For example, a previous study found that LightGBM produced an accuracy of 80%, while Random Forest produced an accuracy of 80%. Chang et al. used XGBoost to obtain 90% accuracy and 94% AUC, while Jumaa et al. used an MLP to report 93% accuracy. By comparison, our suggested stacking ensemble model obtained 93.69% accuracy, 95.59% precision, 95.55% recall, and 97.81% AUC.

The methodical and data-driven approach used at every step of the model development process is principally responsible for this notable increase. In particular, this improved predictive capability was made possible by the careful pre-processing of the data, the discovery and use of the best class balancing method (SMOTE+ENN), the strategic feature selection using RFECV, and the clever integration of various base learners using the stacking ensemble approach.

The model was specifically adjusted to minimize false negatives, which is a crucial need in financial risk assessment because the penalty of missing a defaulter is exceptionally large. This was made possible by the process's concentration on recall as a major optimization indicator [46].

## 5.3. Implications, Limitations, and Future Work

The study's conclusions have important ramifications for peer-to-peer lending platforms and financial institutions. A reliable and extremely precise method for forecasting loan failure risk is provided by the created stacking ensemble model, which can immediately lower financial losses and increase profitability. A reproducible framework that can be modified for different credit risk assessment scenarios outside of the LendingClub dataset is provided by the methodical approach to data pre-processing, class imbalance handling, and model tuning. Additionally, a deeper comprehension of the factors that contribute to default is made possible by the model's explainable character, which is made possible by SHAP values. This promotes increased transparency and confidence in automated lending decisions [47].

This study admits its shortcomings despite its noteworthy contributions. The biggest obstacle was data availability because there is still limited access to sizable, current, and openly accessible financial datasets for predicting credit risk. A more recent dataset might provide insights into current borrower habits and economic trends, even though the LendingClub dataset was a great source of information. Furthermore, the computational complexity of certain methods and models, especially during ensemble training and hyperparameter tuning, made it necessary to carefully employ XGBoost as a test model for first assessments. The amount of thorough testing and improvement that could be done was also limited by time.

However, these drawbacks do not lessen the significance of the study; rather, they point to interesting directions for further investigation. Future research may include:

- **Model-Specific Pre-processing Optimization:** Future studies could examine whether various pre-processing and normalization techniques are more successful when customized for particular individual models within the ensemble, as opposed to relying solely on a single test model (XGBoost) to choose the best pre-

processing strategies. This might open the door to more performance improvements.

- **Generalizability Across Diverse Datasets:** The suggested approach and framework's robustness and generalizability would be further confirmed by applying it to additional credit datasets from various financial markets or geographical areas. This would make it easier to determine how well the stacking ensemble approach adapts to different data types and economic situations.

- **Advanced Explainable AI Techniques:** Beyond SHAP, additional research into advanced Explainable AI (XAI) methodologies may yield even more insightful model predictions, especially for intricate ensemble systems [48].

# Chapter 5

## Conclusion

This paper aimed at enhancing the accuracy of credit default risk forecast in financial institutions in an effort to limits financial losses and maintain the stability of the lending operations of financial institutions. The research was methodological in its approachs considering thoughtful evaluatios and integration of various methods along the predictive pipeline, which include using advanced ensembled algorithms and balancing out data imbalance.

Among its key results is a successful desigdn and confirmation of a powerful stacking ensemble framework, which demonstrated impressive potentials to identify actual defaulters through an impressive recall rating of 95.5% to exactly predict default risk. The researchers also systematically calculated the optimal practices at each stage of model development, such as the SMOTE + ENN approach to the utilizations of class imbalance, the so-called robust scaler to preserved data normalization levels, and the so-called winsorize approach to the presence of outliers [49]. The use of SHapley Additive exPlanations (SHAPs) to enable an explainable model is another contribution that provides insight into the feature contributions as required by regulatory compliance and practical understanding.. Important discoveries highlight how successful the suggested approach is. SMOTE + ENN shown to be the most successful in managing the significant class imbalance of the LendingClub dataset, with balanced datasets continuously outperforming imbalanced ones. It was shown that the best pre- processing approach prioritized recall by using 'robust scaler' for normalization and 'winsorize' for outliers. Interest rate, credit score (FICO), and loan period were the most significant predictors of default, according to feature importance analysis. The stacking ensemble model, especially "Stacking A" (which combines all base learners), produced better overall results with an accuracy of 93.69%, precision of 95.59%, recall of 95.55%, and an AUC of 97.81%, even though XGBoost had outstanding individual performance.

This steady progress demonstrates the value of astute cooperation between many models.

This study has significant practical ramifications. Financial institutions have a reliable instrument to lower default rates and losses thanks to the suggested ensemble model. Beyond the LendingClub dataset, the methodical framework is quite adaptable and may

be modified for different credit risk assessment scenarios. Furthermore, the model's explainable nature—made possible by SHAP values—offers vital transparency, enabling institutions to comprehend the factors influencing forecasts and improve lending practices. Notwithstanding these developments, the study had drawbacks, chief among them being the lack of sizable, current, and openly available financial databases, which limited the amount of data that could be gathered. Furthermore, the amount of thorough testing and improvement was impacted by time constraints as well as the computational complexity of several methods and models [50].

However, these restrictions open up new avenues for investigation. To further improve predictive accuracy and adaptability, it is advised to investigate model-specific optimization for pre-processing techniques, validate the methodology on a variety of credit datasets from various financial markets, and look into more sophisticated ensemble strategies or dynamic ensemble selection techniques.

# REFERENCES


[1] J. Crook, D. Edelman, and L. Thomas, "Recent developments in consumer credit risk assessment," *Eur. J. Oper. Res.*, vol. 183, no. 3, pp. 1447–1465, Dec. 2007.

[2] T. Bellotti and J. Crook, "Support vector machines for credit scoring and discovery of significant features," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3302–3308, Mar. 2009.

[3] L. Lessmann, B. Baesens, H. Seow, and L. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," *Eur. J. Oper. Res.*, vol. 247, no. 1, pp. 124–136, Nov. 2015.

[4] A. Brown and J. Mues, "An experimental comparison of classification algorithms for imbalanced credit scoring data sets," *Expert Syst. Appl.*, vol. 39, no. 3, pp. 3446–3453, Feb. 2012.

[5] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 785–794.

[6] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[7] J. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001.

[8] Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, 1997.

[9] S. Haykin, *Neural Networks and Learning Machines*, 3rd ed. Prentice Hall, 2009.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 1135–1144.

[11] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 4765–4774.

[12] N. V. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[13] H. He and E. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009.

[14] G. Batista, R. Prati, and M. Monard, "A study of the behavior of several methods for balancing machine learning training data," *SIGKDD Explor.*, vol. 6, no. 1, pp. 20–29, 2004.

[15] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, 1992.

[16] L. Breiman, "Stacked regressions," *Mach. Learn.*, vol. 24, no. 1, pp. 49–64, 1996.

[17] S. Raschka, "Stacking classifiers for credit risk modeling," *Mach. Learn. J.*, vol. 5, pp. 12–25, 2020.

[18] S. B. Kotsiantis, "Decision trees: A recent overview," *Artif. Intell. Rev.*, vol. 39, pp. 261–283, 2013.

[19] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*, 2nd ed.