NOVA IMS Information Management School

MEGI

Master Degree Program in

Statistics and Information Management

Optimizing Credit Scoring Models in Face of Global Economic Uncertainty: A Comprehensive Risk Analysis in Banking loans

David Manuel Pereira Susana

Master Thesis

presented as partial requirement for obtaining the Master Degree in Statistics and Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Optimizing Credit Scoring Models in Face of Global Economic Uncertainty: A Comprehensive Analysis of Profitability and Risk in Banking loans

by

David Manuel Pereira Susana

Master Thesis presented as partial requirement for obtaining the Master's degree in Statistics and Information Management, with a specialization in Risk Analysis and Management

Supervisor: Professor Dr. Jorge Bravo

Julho, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisboa, Novembro 2024

ACKNOWLEDGEMENTS

First of all, I'd like to thank my bosses at work, as they were always available to help me with the timetable for the conclusion of this thesis. I would also like to thank my parents and my girlfriend for all the support they have given me over the last two years of my master's degree. Without them, I wouldn't have been able to reconcile my professional life with my academic life. Last but not least, I would like to thank Professor Jorge Bravo for all his support and willingness to help me complete this thesis.

ABSTRACT

In the contemporary financial landscape, influenced by various global events, banks have become increasingly cautious in extending credit, both for housing and for other consumer purposes. In an increasingly uncertain economic world, banks must develop ever more robust and effective models. This topic is of significant importance to explore and update, as banks directly impact a substantial portion of the global population through housing and consumer credit. This thesis aims to contribute to the development of models with current data, and consequently, assist the population. In this research, Python programming language will be employed to utilise a Machine Learning Approach for credit scoring analysis. Methods to be utilised include: Logistic Regression; Random Forest; Gradient Boosting; XGBoost. To determine the best model, the evaluation will be performed on four metrics: Accuracy; AUC Score; Type I Error; Type II Error. The XGBoost method was the best performer on all evaluated metrics. In the course of reviewing the selected literature, previous work were found that explored this subject solely with the objective of identifying the best Machine Learning method to create the optimal model for determining customer defaults. However, several questions emerged: 'Will machine learning models be better than Logistic Regression?' 'Is the model that accepts more credit the safest for the bank, considering the Profit / Risk ratio?' This thesis aims to answer these questions and determine not only the best model for the bank, but also its profits.

KEYWORDS

Analytical Models; Banking; Machine Learning; Credit Risk; Credit Scoring; Loan Defaults

Sustainable Development Goals (SDG):

INDEX

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Banks hold an important role in our society as they serve as vital financial intermediaries, facilitating capital formation, investment, and economic growth, providing essential payment and settlement services, managing risks, extending credit, implementing monetary policies, promoting financial inclusion, offering savings and wealth management opportunities, contributing to currency issuance, maintaining financial stability, managing crises, generating employment, and overall, playing a central role in fostering the economic well-being and prosperity of individuals and businesses.

Within this multifaceted landscape, the issue of loan default emerges as a significant consideration, transcending mere financial implications to exert a noteworthy influence on the broader economic arena. Particularly in the traditional banking industry, where loans are intricately linked to economic activities, defaults by borrowers can trigger a ripple effect, leading to a reduction in credit availability. This, in turn, has repercussions on businesses and individuals, influencing economic growth and stability, thus emphasizing the intricate connection between banking practices and the overall health of the economy.

Moreover, the dimension of investor confidence emerges as a critical factor affected by the occurrence of loan defaults. In the traditional banking sector, defaults have the potential to undermine investor trust, resulting in a reduction in deposits and heightened borrowing costs. This phenomenon is mirrored in the Internet financial sector, where investors providing capital for loans can experience erosion in confidence due to defaults, impacting stock values and shaping perceptions of platform stability. Understanding and addressing these critical dimensions are imperative for ensuring the sustained trust and support of investors, further underscoring the integral role of banks in maintaining financial equilibrium and fostering a resilient economic landscape.

Nowadays, in advanced economies, a high proportion of loans are automatically decided using frameworks where the credit score is the central, if not the unique, indicator of the borrowers' credit risk. Within the OECD countries, Probability of Default models, abbreviated as PD models, play a fundamental role in computing the regulatory capital of banks under the Internal Ratings-Based (IRB) approach outlined in the revised Basel Framework. Specifically, for retail exposures, the assessment of the borrower's PD, coupled with the Loss Given Default (LGD) and the Exposure at Default (EAD), serves as crucial input risk parameters for the calculation of minimum capital requirements.

According to (Sheikh et al., 2020), while retail banking institutions offer a diverse range of services, their primary revenue stream originates from the loans they extend. This is due to the application of interest rates on each loan, resulting in profits for the bank. Nonetheless, credit lending poses significant risks, with the challenge lying in the ability to differentiate between creditworthy applicants, those unlikely to default, and unworthy applicants who may struggle to fulfill the contract and repay the loan.

For banks to be able to weigh the risk of their prospective borrower being able to fulfill their repayments, they collect significant information both on the borrower and the underlying property of the mortgage. The outcome of these gathered data is referred to as Credit Scoring. Credit scoring involves the use of analytical methods to transform relevant data into numerical measures that inform and determine credit decisions (Ashofteh & Bravo, 2021) and is a concept that emerged about 70 years ago with (Durand, 1941), which indicates the creditworthiness of loan applicants. These applicants are then ranked according to their credit score for the determination of their default probability and the subsequent classification into either non-defaulter applicants or defaulters. Credit scoring involves the use of analytical methods to transform relevant data into numerical measures that inform and determine credit decisions (Ashofteh & Bravo, 2019, 2021).

Over the years, machine learning models have been considered important tools for building predictive models. An important research stream in the academic literature is related to the development of credit scoring models using machine learning classifiers. However, building an optimum credit score prediction model is a potential area of research. To build a robust, accurate, and sensitive machine learning prediction model, the information of input predictors is very important. Feature selection refers to the use of methods to evaluate the informative features and reduction of data dimensionality. In the literature, many feature selection techniques have been tested and showed improvement in credit score prediction (Trivedi, 2020). By combining the optimal model for classifying clients as defaults or non-defaults with the most effective estimators to determine internal IRB measures, we aim to maximize the profit-to-risk ratio for the bank.

MOTIVATION

In the contemporary financial landscape, influenced by various global events (Dornigg, 2022), banks have become increasingly cautious in extending credit, both for housing and other consumer purposes.

In June 2004, the Basel Committee on Banking Supervision issued a revised framework on International Convergence of Capital Measurement and Capital Standards (hereafter “Basel II” or the “revised Framework”). When following the “internal ratings-based” (IRB) approach to Basel II, banking institutions will be allowed to use their internal measures for key drivers of credit risk as primary inputs to their minimum regulatory capital calculation, subject to meeting certain conditions and to explicit supervisory approval. In light of the need under Basel II for banks and their supervisors to assess the soundness and appropriateness of internal credit risk measurement and management systems, the development of methodologies for validating external and internal rating systems is an important issue. More specifically, there is a need to develop means for validating the systems used to generate the parameters (such as PD, LGD, EAD and the underlying risk ratings) that serve as inputs to the

IRB approach to credit risk. In this context, validation comprises a range of approaches and tools used to assess the soundness of these elements of IRB systems.

Several research studies conducted recently have demonstrated that artificial intelligence methods produce better results than traditional statistical models, which are still the most often used method in the field. In particular, Li and Zhou (2012) point out that ensemble methods are considered to have superior performance compared to a range of alternative machine learning strategies. The current financial crisis emphasizes how crucial it is to develop a robust risk management culture. Lending institutions must include strict and reliable credit evaluation procedures in their systems if they want a thriving economy. Effective prediction of credit risk well in advance enables proactive measures to prevent loan defaults, thereby averting the potential onset of a catastrophic recession.

STUDY OBJECTIVES

The banking sector is one of the earliest applications of Machine Learning. Financial institutions meet bad debts and losses every year. In recent years the banking sector has been improving its credit risk management practices through customer profiling, past expenditures, customer transaction behavior, etc. Data Science is a combination of various statistical tools, algorithms and machine learning techniques to extract the hidden patterns from the data which helps to turn into insights. Nowadays, financial organizations take advantage of data science applications to study individual customers' banking profiles and provide the appropriate services by segmenting the customers based on their credit history (Maheswari & Narayana, 2020).

The main purpose of the present thesis is to investigate which supervised machine learning classifiers perform the best at predicting customer loan defaults. The following research question will be addressed:

"Will machine learning models be better than the traditional method (Logistic Regression)? "Is the model that accepts more credit the safest for the bank?"

In this research, Python programming language will be employed to utilize a Machine Learning Approach for credit scoring analysis (Trivedi, 2020; K. Wang et al., 2022). The methods to be utilized include: Logistic Regression (Chamboko & Bravo, 2016, 2019, 2020; Dumitrescu et al., 2022); Random Forest (Dumitrescu et al., 2022; Li & Zhong, 2012; Maheswari & Narayana, 2020); Gradient Boosting (Bentéjac et al., 2021) and XGBoost (Chen & Guestrin, 2016; Liu et al., 2022; Nalluri et al., 2020). The data is publicly available on the website of Freddie Mac.

STUDY RELEVANCE

The expected value of this research is to provide a more economic perspective on the development of credit scoring models. This topic is of significant importance to explore and update, as banks directly impact a substantial portion of the global population through housing and consumer credit. The social impact of these models extends to the reduction of systemic risks. By identifying and managing potential defaults early on, the financial system is better equipped to handle challenges without succumbing to widespread distress. This has a cascading effect on society, preventing the ripple effects of financial crises that can lead to increased unemployment, poverty, and social instability.

In essence, having reliable models for predicting loan defaults is not merely a technical consideration for financial institutions. It goes beyond the realm of profit and loss statements, shaping the economic landscape, preserving jobs, fostering confidence, encouraging responsible practices, and ultimately contributing to a more equitable and stable society.

THESIS OUTLINE

- **Introduction:** The introduction provides the background and context for the study, emphasizing the significance of credit risk modeling in the financial industry;
- **Literature Review:** Provides a comprehensive overview of the existing research and theoretical framework relevant to the study;
- **Methodology:** The methodology chapter describes the research design and approach, detailing the procedures followed in the study;
- **Data Properties and Data Preprocessing:** This chapter delves into the characteristics of the dataset, including its sources, structure, and key variables;
- **Models Preparation:** In this section, each model (Logistic Regression, Random Forest, Gradient Boosting, XGBoost) is described in detail;
- **Performance Measures:** This chapter defines and explains the importance of key performance metrics such as accuracy, AUC, Type I error, and Type II error;
- **Results and Discussion**: The results and discussion chapter presents the performance results of each model based on the selected metrics;
- **Conclusion:** The conclusion summarizes the key findings of the study, highlighting the main contributions to the field of credit risk modeling;
- **Limitations and Recommendations:** This chapter addresses the limitations of the study. Provides recommendations too for future research, suggesting areas for further exploration.

2. LITERATURE REVIEW

This section furnishes essential context on credit risk and machine learning, along with a summary of past research works in credit scoring that have utilized machine learning methodologies.

2.1. CREDIT RISK

Credit risk can be defined as the potential financial loss that a lender may incur due to the failure of a borrower to fulfill their repayment obligations according to the agreed terms and conditions. It encompasses the uncertainty and likelihood of default by the borrower, reflecting the possibility that the borrower may be unable or unwilling to meet their contractual obligations, leading to adverse consequences for the lender. Assessing and managing credit risk is crucial for financial institutions to make informed lending decisions and maintain a sound and sustainable financial position.

Given the requirements stipulated by Basel II for banks and their supervisors to evaluate the effectiveness and suitability of internal systems for measuring and managing credit risk, the validation of both external and internal rating systems emerges as a significant concern. This pertains particularly to establishing mechanisms for validating the systems responsible for generating parameters like PD, LGD, EAD, and the foundational risk ratings. These parameters play a crucial role as inputs in the Internal Ratings-Based (IRB) approach to credit risk (Engelmann & Rauhmeier, 2006):

Banks calculate Probability of Default (PD), Exposure at Default (EAD) and Loss Given Default (LGD) parameters as part of their credit risk management processes. These parameters are essential components in estimating the potential credit losses associated with their loan portfolios.

These parameters are critical for several reasons:

- Risk Management: PD, EAD and LGD are integral to assessing credit risk within a bank's portfolio. By understanding the potential exposure and expected losses in the event of default, banks can manage their risk more effectively;

- Capital Adequacy: Regulatory frameworks, such as Basel III, require banks to hold sufficient capital to cover potential losses. Accurate calculation of EAD and LGD helps in determining the amount of economic capital that should be set aside for credit risk;
- Credit Pricing and Decision-Making: PD, EAD and LGD influence credit pricing and lending decisions. Banks need to appropriately price their loans based on the level of risk involved. Accurate estimation of potential losses also helps in making informed decisions regarding the approval or rejection of credit applications;

- Financial Reporting: Banks are required to disclose their risk exposures, PD, EAD and LGD contribute to the assessment and reporting of credit risk in financial statements.

2.2 CREDIT SCORING

Credit scoring can be formally defined as a statistical (or quantitative) method that is used to predict the probability that a loan applicant or existing borrower will default or become delinquent (Mester, 1997). This helps to determine whether credit should be granted to a borrower. Credit scoring can also be defined as a systematic method for evaluating credit risk that provides a consistent analysis of the factors that have been determined to cause or affect the level of risk (Fensterstock, 2005). The objective of credit scoring is to help credit providers quantify and manage the financial risk involved in providing credit so that they can make better lending decisions quickly and more objectively (Koh et al., 2006).

The methods and techniques of Credit scoring are used by credit-granting institutions in determining whether to approve or deny credit to an applicant. The primary objective is to assess the likelihood that the applicant will default. If this probability is low, the applicant is deemed a good customer, indicating a higher probability of loan repayment. Conversely, if the probability is high, the applicant is labeled a bad customer, suggesting a greater likelihood of undesirable financial behavior in the future. Credit scoring is closely related to the probability of default (PD) of a loan. Credit scoring models assign a numerical score to each borrower based on these factors. The score is then used to categorize individuals into different risk categories. A higher credit score indicates a lower probability of default, while a lower score suggests a higher likelihood of default.

Lenders use these credit scores, along with other risk assessment tools, to make informed decisions about whether to approve a loan, what interest rate to charge, and what credit limits to set. The probability of default is a crucial component in these decisions, as it helps lenders manage and mitigate the risk associated with lending money.

2.2.1 SCORING FORMULATION

A credit scoring model is a simplification of the reality. The output is a prediction of a given entity, actual or potential borrower, entering in default in a given future period. When applied to the assessment of credit risk, we are essentially addressing a supervised learning problem where the goal is to predict the default status (good, bad) based on a set of input characteristics.

The objective of supervised learning classification methods is to identify a function that effectively separates individuals into the "good" and "bad" classes within the problem space. A robust model is crucial for accurately distinguishing between the two classes by capturing

sufficient information to predict the probability of default (belonging to the "bad" class) based on known occurrences of default in the past.

2.2.2 CREDIT SCORING MODELS

Credit-scoring model is a good and effective tool for global financial institutions. Last few years, numerous credit-scoring models have been proposed in literatures to evaluate the consumer loans and improve the credit-scoring accuracy (Crook et al., 2007).

Linear regression methods (Orgler, 1970) have become an essential component of any data analysis concerned with describing the relationship between a response variable and one or more independent variables. Linear regression has been used in credit scoring applications, as the two-class problem can be represented using a dummy variable. Using a Poisson regression model instead could be used to accommodate cases where the customer makes varying degrees of partial repayments. Variables can be analyzed by credit analysts with linear regression to set up a score for each factor and then to compare this with the bank's cut-off score (if a customer's score passes the score cut-off, the credit will be granted).

With the rapid advancement of machine learning in the past decades, AI-based methods, which extensively use machine learning algorithms in data processing and modeling, have been introduced to the credit scoring domain. Machine learning algorithms mainly include decision tree (DT), support vector machine (SVM), artificial neural network (ANN), and evolutionary computation techniques. Some experimental studies have demonstrated the advantages of AI-based methods relative to statistical ones in developing scoring Models (Xia et al., 2020). A recent strand of literature uses model combinations (ensembles) of heterogeneous models aggregated using, for instance, a metalearning model or adopting Bayesian model ensembles (Bravo, 2020; Bravo & Ashofteh, 2023; Bravo & Ayuso, 2021; Raimundo & Bravo, 2023).

2.2.3 CREDIT SCORING APPLICATIONS

In the early years, financial institutions used credit scoring mainly to make credit decisions for loan applications. Over the past 25 years, however, the application of credit scoring has grown from making credit decisions to making decisions related to housing, insurance, basic utility services, and even employment (Koh et al., 2006). Some examples are:

- Credit decisions for loan applications;
- To help set credit limits, manage existing accounts, and forecast the profitability of consumers and customers;

- Use credit scores as a decision support tool to identify their target market for credit cards;

- Credit scores are also used as a basis to adjust premiums.

Applications of credit scoring in housing-related choices seek to evaluate the risk involved in lending or leasing and support lenders and landlords in making just and knowledgeable choices. People should keep up strong credit habits, like don´t spend a lot of money in casinos (it may decrease your credit score) , in order to increase their chances of getting favorable conditions while making financial decisions relating to housing or other type of credit.

2.3 Machine Learning

Machine Learning is the development of algorithms and models that enable computers to learn from and make predictions or decisions based on data. Instead of being explicitly programmed to perform a task, a machine learning system uses statistical techniques to learn patterns and relationships within the data, allowing it to generalize and make predictions or decisions on new, unseen data. The key idea behind machine learning is to enable computers to improve their performance on a specific task over time by learning from experience or data, without being explicitly programmed for that task.

As said by (Nariya et al., 2023), effectively evaluating the performance of predictive computational models is a crucial aspect of machine learning, so in recent years, machine learning has gained the interest of researchers as they are trying to implement models and algorithms to perform various important tasks and facilitate everyday life. Currently, machine learning is the IT domain that contributes the most to business forecasting problems according to (Ngai et al., 2011).

2.3.1 Machine Learning Models and Methods

Machine learning models have shown significant promise in enhancing credit risk assessment by identifying complex, non-linear patterns in data. This is especially true for models that utilize ensemble learning techniques. Ensemble models often outperform single classification algorithms in terms of prediction accuracy and generalization. As a result, many modern credit risk modeling strategies now rely heavily on ensemble learning. The core concept of ensemble methods is to create several individual models and then combine their outputs to achieve more accurate predictions than a single model could provide. Ensemble modeling has rapidly become a crucial tool, particularly in the field of credit risk. Ensemble learning is a machine learning paradigm where multiple learners are trained to solve the same problem (Polikar, 2006). In contrast to ordinary machine learning approaches that try to learn one hypothesis

from the training data, ensemble methods try to construct a set of hypotheses and combine them to use (Zhou, 2012).

It's important to recognize that non-parametric models can be integrated into ensemble models; they are not mutually exclusive. Some ensemble techniques actually harness the strengths of non-parametric models. For instance, a decision tree, which is not an ensemble method by itself, can serve as the base model for ensemble techniques. Methods like Random Forests and Gradient Boosting Machines combine the predictions of multiple decision trees to produce a more robust and accurate model.

2.4 Related Works using Machine Learning to predict Loan Default

While the incorporation of machine learning in the field of Finance is a relatively recent concept, extensive research has been undertaken in this domain. (Khandani et al., 2010), (Butaru et al., 2016), (Fitzpatrick & Mues, 2016) and (Sealand, 2018) employed machine learning in predicting loan default. Some of these studies used small datasets with several thousand mortgages, while other used dataset of millions of mortgages.

This paper (G. Wang et al., 2011) conducted a comparative assessment of the performance of three popular ensemble methods, i.e., Bagging, Boosting, and Stacking, based on four base learners, i.e., Logistic Regression Analysis (LRA), Decision Tree (DT), Artificial Neural Network (ANN) and Support Vector Machine (SVM). Experimental results reveal that the three ensemble methods can substantially improve individual base learners. In particular, Bagging performs better than Boosting across all credit datasets. Stacking and Bagging DT in our experiments, get the best performance in terms of average accuracy, type I error and type II error.

By the year 2016, (Ala'raj & Abbod, 2016) did a paper where the main objective was improve accuracy by exploring a new combination method in the field of credit scoring by developing a new combination rule whereby the ensemble classifiers can work and collaborate as a group or a team in which their decisions are shared between classifiers.

In (Mamonov & Benbunan-Fich, 2017) study, were examined the data that were available to Fannie Mae prior to mortgage origination (ex-ante) and they apply data mining techniques to explore the systematic relations that were present at the mortgage origination stage that may yield clues to mortgage risks. The methods they used were Logistic Regression, Decision tree, Random Forest, Boosted trees, SVM, Neural network and the ANN (Neural network) was the best performing model in their analysis.

In (Addo et al., 2018) investigation, he built a binary classifiers based on machine and deep learning models on real data in predicting loan default probability. The top 10 important features from these models are selected and then used in the modeling process to test the stability of binary classifiers by comparing their performance on separate data. He concluded that algorithms based on artificial neural networks (ANN) do not necessarily provide the best

performance and that regulators need also to ensure the transparency of decision algorithms to avoid discrimination in and a possible negative impact on the industry.

This paper discussed how data science can impact the banking sector to improve their analysis of identifying risk by preprocessing the historical data of customers and building the model using machine learning techniques (Maheswari & Narayana, 2020). Logistic regression with SGD training result in better predictions than the others.

This article introduces penalized logistic tree regression (PLTR) with predictive variables given by easy-to-interpret endogenous univariate and bivariate threshold effects. These effects are quantified by dummy variables associated with leaf nodes of short-depth decision trees built with singletons and couples of the original predictive variables (Dumitrescu et al., 2022). They show that PLTR performs better in out-of-sample than traditional linear and non-linear logistic regression, while being competitive relative to the random forest method and leads to a significant reduction in misclassification costs.

2.4.1 Results from Related Works

The results from related previous work are summarized in the following table.

Table 1 *Results from related works*

Source: Authors preparation

Authors	Models	Measures
(G. Wang et al., 2011)	Stacking; SVM; ANN; Decision Tree; Logistic Regression; Boosting; Bagging;	Accuracy; Type I Error; Type II Error
(Ala'raj & Abbod, 2016)	Logistic Regression; MARS; NN; SVM; DT; RF; NB;	Accuracy; AUC; H-measures; Brier Score
(Mamonov & Benbunan-Fich, 2017)	Logistic Regression; Decision Tree; Random Forest;	Accuracy; Type I Error; Type II Error

	Boosted Trees; SVM; Neural Network	
(Addo et al., 2018)	Logistic Regression; Random Forest; Boosting Approach	Accuracy; AUC
(Maheswari & Narayana, 2020)	Logistic Regression; Random Forest; KNN	Accuracy; Precision; Recall
(Dumitrescu et al., 2022)	Logistic Regression; Random Forest; SVM; NN; PLTR	Brier Score; KS; PCC; AUC; PGI

3. METHODOLOGY

The methodology chapter is a guidebook that explains how the research questions and goals are tackled. It's all about showing how the data is collected, analyzed, and understood, making sure the findings are trustworthy and reliable.

The field of data mining has grown and become more established. Both academics and industry professionals are working on setting standards in this area. In this thesis, we've chosen to focus on SEMMA (Sample, Explore, Modify, Model, and Assess) and CRISP-DM because they are widely recognized as the most popular methodologies for modeling large datasets.

SEMMA (Lin & McClean, 2001) was developed by SAS Institute and is commonly used in data mining and predictive analytics projects. Here's how SEMMA can be applied to credit scoring. This methodology offers an easy-to-understand process, allowing an organized and adequate development and
maintenance of projects. It thus confers a structure for its conception, creation and evolution, helping to present solutions to business problems as well as to find business goals.

- **Sample**: Obtain a representative sample of data for analysis, including customer information, credit history, financial data. Ensure that the sample adequately represents the population of interest for credit scoring;

- **Explore**: Explore the data to understand its characteristics, distributions, relationships, and potential patterns. This stage consists on the exploration of the data by searching for unanticipated trends and anomalies;

- **Modify**: Preprocess and modify the data to prepare it for modeling. This may involve handling missing values, outlier detection and treatment, feature engineering, and transforming variables as needed for modeling;

- **Model**: Select and build predictive models for credit scoring using techniques such as logistic regression, decision trees, neural networks, or ensemble methods. Train the models on the prepared data and tune their parameters to optimize performance;

- **Assess**: Evaluate the performance of the credit scoring models using validation data or cross-validation techniques. Assess the models' predictive accuracy, discrimination ability, calibration, and stability over time. Compare different models and select the best-performing one for deployment.

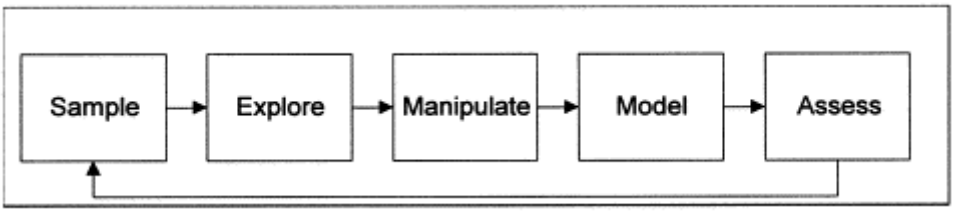


Figure 1 *SEMMA Process*

Source: *A data mining approach to the prediction of corporate failure* (Lin & McClean, 2001)

The CRISP-DM (Azevedo & Santos, 2008) process provides a structured and iterative approach to data mining projects, allowing teams to effectively manage and execute complex analytics tasks. It emphasizes the importance of understanding the business context, exploring the data, preparing it properly, building and evaluating models, and deploying them into operational use.

- **Business Understanding:** First of all, we must acknowledge the project's business objectives. It involves knowing what these goals are and then converting this to defining a data problem clearly and proposing how we can solve it.

- **Data Understanding:** After that, we collect some initial data and become acquainted with them. This entails identifying any problems, discovering early insights, as well as singling out interesting pieces of data that could potentially help us to uncover obscured trends.

- **Data Preparation:** At this stage, raw data is cleaned up and organized in preparation for analysis.

- **Modeling:** In this step, various methods are employed in building models while optimizing their settings for better performance.

- **Evaluation:** We now assess the accuracy of our predictions based on model evaluation metrics against the business objectives set earlier.

- **Deployment**: Finally, the best model is put into practice and integrated with the current systems. The performance of this model should be monitored continually so as to make any adjustments if necessary.

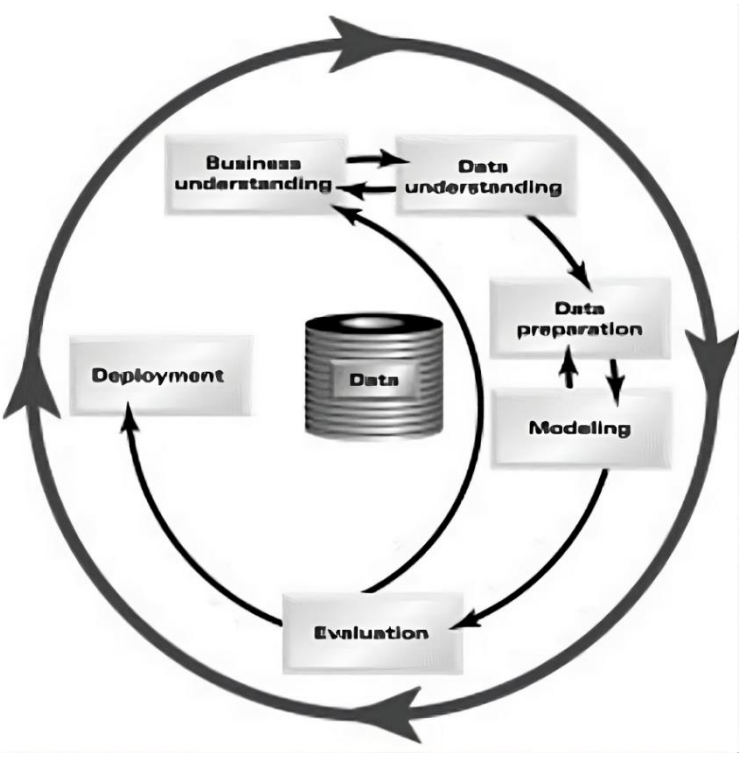


Figure 2 *CRISP-DM Process*

Source: *KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW* (Azevedo & Santos, 2008)

CRISP-DM will be the chosen methodology, because it covers the entire data mining process, including understanding the business problem, data preparation, modeling, evaluation, and deployment. This methodology emphasizes understanding the business context and objectives before diving into data analysis. This is particularly relevant for credit scoring, where the models need to align closely with the goals and requirements of financial institutions.

4. MODELS PRESENTATION

In this section, we will briefly explore the prediction models used in this thesis. The models investigated include Logistic Regression, Random Forest, Gradient Boosting and XGBoost.

4.1 LOGISTIC REGRESSION

Logistic regression is a type of regression analysis that is good for binary outcome prediction. When used to estimate the probability of default, given various predictor variables, logistic regression analysis is well-suited to the development of credit scoring. This is highly important as a technique for financial institutions in making informed decisions regarding the approval of loans, decisions on interest rates, and credit limits while at the same time managing the associated risk that comes with lending. As with linear regression, we are interested in understanding the relationship between a dependent variable and one or more independent variables. But, in this case, we want to predict a categorical output variable y, which has just two outcomes, a binary output variable, rather than a continuous output variable.

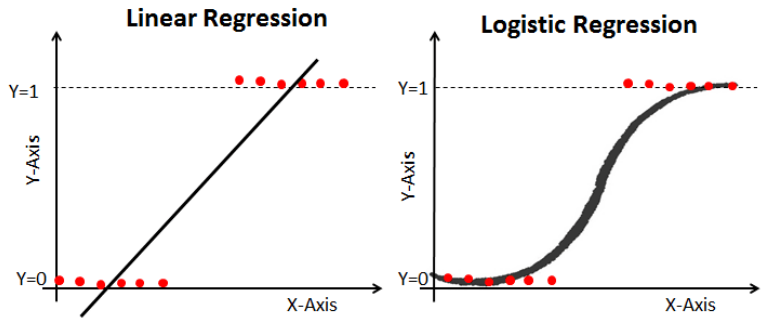


Figure 3 *Shape of Linear and Logistic Regression*

Source: ResearchGate

The logistic function is given by formula:

$$f(x) = \frac{1}{1+e^{-(\beta 0+\beta 1x)}} \qquad (1)$$

where $f(x)$, in this scope, represents the probability of an output variable, $\beta 0$ is the linear regression intercept and $\beta 1$ is the multiplication of the regression coefficient by x value of the independent variable. In the context of the present thesis we are dealing with binary logistic regression, in this approach, the dependent variable has a dichotomous nature, i.e. it has only two possible outcomes (default or no default).

In this case, if a client is predicted to default the response variable takes a value equal to one; if a client is predicted to not default, then the response variable takes a value equal to zero. This is represented below in the form of a logistic equation.

$$P = P(Client\ in\ Default\ is\ 1) = \frac{1}{1+e^{-(\beta 0+\ \beta 1x+\cdots+\beta kxk)}} \quad (2)$$

where k is the number of independent (explanatory) variables. Therefore, the logistic regression formula for default loans becomes:

$$\frac{f(x)}{1-f(x)} = e^{-(\beta 0+\ \beta 1x+\cdots+\beta kxk)} \quad (3)$$

where $f(x)$ is the probability of the loan being of the nature default.

4.2 DECISION TREE

Imagine a situation where a dataset describes credit applicants with multiple features or characteristics, represented by the notations $\mathrm{x1}, \mathrm{x2}, \ldots, \mathrm{xn}$. There are two types of applicants: those with "good credits" and those with "bad credits". Creating a classifier that successfully separates these two groups is the aim of a credit scoring model.

By using binary splits to divide the data iteratively, a decision tree method works. A root node is initially made up of a mix of candidates with and without good credit. In order to distribute primarily good credits on one side and predominantly bad credits on the other, the algorithm then assesses several binary splits in order to identify the attribute x and associated cutoff value c that give the optimal separation.

Consider Figure 4, where dividing the root node into instances with attribute xi greater than or equal to c and those with xi less than *c* yields the best result. Until a predetermined stopping requirement is satisfied, this process is repeated for each additional daughter node.

The splitting property and cutoff value are chosen based on a node's purity p, which is the percentage of good credit occurrences within it. The objective is to reduce the total of the daughter nodes' Gini indices, p(1−p). No more splitting takes place if the sum of the Gini indices of the daughter nodes for a specific attribute or cutoff value is greater than that of the parent node. Daughter nodes are more internally homogeneous than parent nodes thanks to the algorithm's minimization of the Gini index, which gauges population variety in a node.

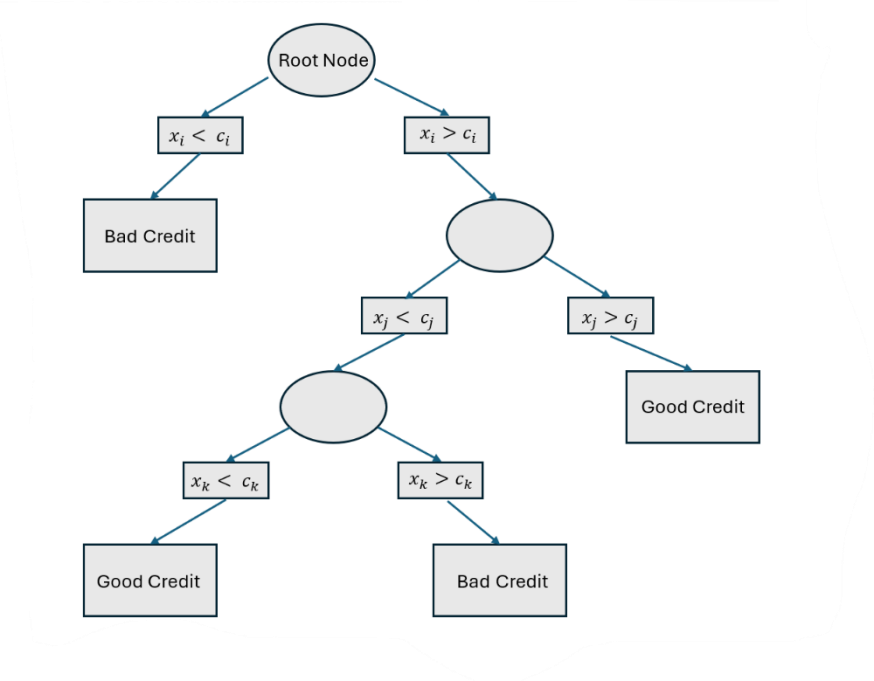


Figure 4 *Example of Decision Tree*

Source: Authors Preparation

Unsplit nodes are denoted by “leafs” and are depicted by rectangles in Figure 4. The leafs are classified according to the most prevalent class in them. A leaf is called “good credit leaf” if it contains several good credit applicants larger than the number of bad credit applicants. Otherwise, it is called “bad credit leaf”. A good (bad) credit is correctly classified if it lands on a good (bad) credit leaf.

Decision trees have been available since 1980 and have been applied to the development of credit scoring models. They are a powerful and flexible classifier. However, a well-known limitation of decision trees is their instability, since small fluctuations in the data sample may result in large variations in the classifications assigned to the instances.

4.3 Ensemble Methods

In machine learning, ensemble methods are applied to regression problems. They integrate multiple independent weak predictors to come up with a more accurate and reliable predictive model. This is done with the aim of exploiting complementary features from the different models to make up for the shortcomings in each individual predictor, in turn improving generalization. In general, an ensemble method combines multiple models predictions to achieve better performance than any single model. Generally, ensemble methods perform incredibly well in synthesizing complex inter-variable correlations to understand and gauge the underlying patterns in the data. This study will employ the ensemble algorithms belonging to the bagging and boosting families.

In general, there are two types of ensemble learning (Mohammed & Kora, 2023): homogeneous (which includes classifiers of the same sort) and heterogeneous (which includes classifiers of different kinds). Boosting and bagging are part of the homogenous integration.

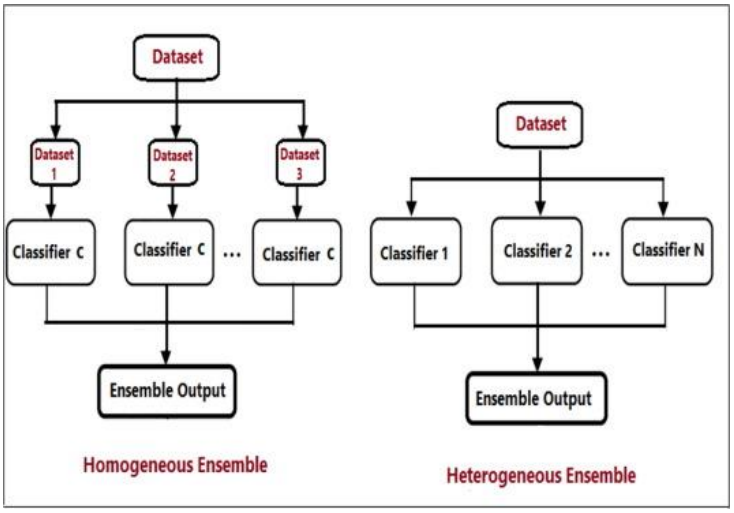


Figure 5 *General Framework of Ensemble*

Source: A comprehensive review on ensemble deep learning: Opportunities and challenges, (Mohammed & kora, 2023)

The basic learner generation procedure generally divides ensemble learning into two approaches:

- Parallel ensemble methods: All base learners are generated simultaneously. The primary driving force behind this approach is the utilization of learner independence.

- Sequential ensemble methods: The generation of base learners occurs in succession. The intention is to leverage the interdependence among students. The learners are built in a sequential fashion to prevent the mistakes made by earlier learners, hence increasing the overall performance.

4.3.1 BAGGING

Bagging is a representative of parallel ensemble methods. Is one of the most popular techniques for constructing ensembles, which is to shape several different training sets with a bootstrap sampling method, then to train base learners on each training set, and the final model will be obtained by aggregating these base learners. These models' predictions are then aggregated, typically by averaging in regression problems. This process helps to mitigate individual model biases and reduce the variance of the overall estimator. By leveraging the

collective insights of multiple predictors, bagging methods enhance the stability and robustness of the predictive model, leading to superior generalization performance on unseen data.

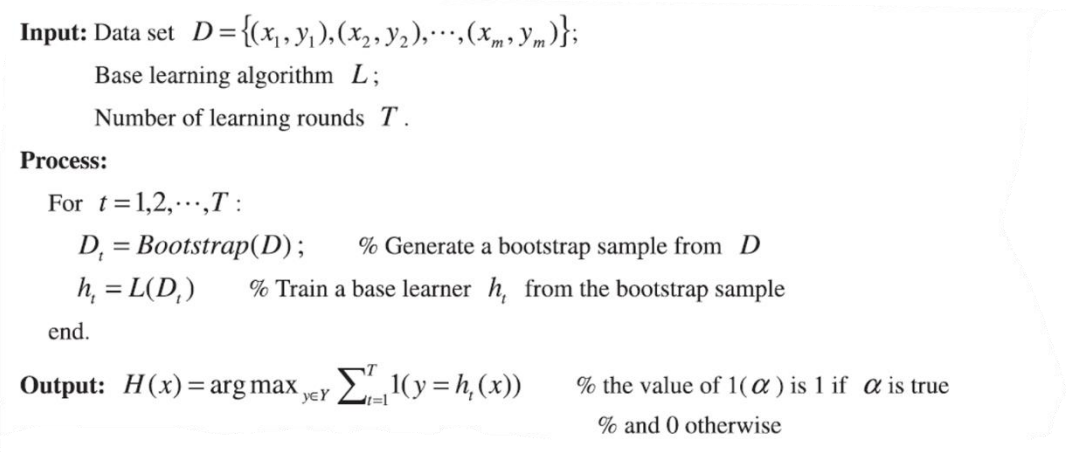


Figure 6 *Bagging Algorithm*

Source: A comparative assessment of ensemble learning for credit scoring, (G. Wang et al., 2011)

Random Forests algorithm is a good example of bagging. There are several challenges to implementing the bagging method: determining the optimal number of base learners and subsets and the maximum number of bootstrap samples per subset. In addition, the determine of fusion method of integrating the outputs of the base classifiers from various voting methods. In summary, the bagging method uses parallel ensemble techniques where baseline learners are generated simultaneously, as there is no data dependency and the fusion methods depend on different voting methods.

4.3.1.1 RANDOM FOREST METHOD

Random Forest is a bagging of decision tree which randomly restrict the features used in each split and operates by creating multiple decision trees in training process (Mhammedi et al., 2023). There is no correlation between each decision trees in an Random Forest, after generating numbers of trees, the final decision class is based on a vote when a new sample comes in and each decision tree in the Random Forest will make a judgment on which category the sample belongs to.

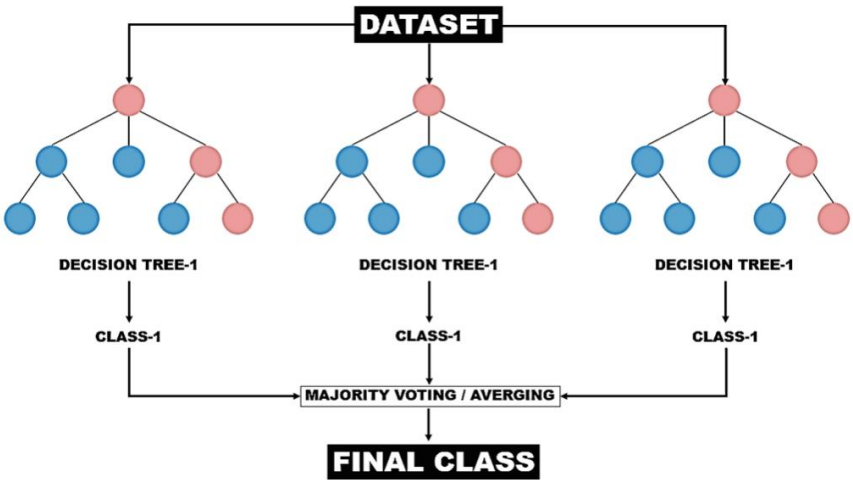


Figure 7 *Random Forest Framework*

Source: ResearchGate, (Mhammedi et al., 2023)

Algorithm of Random Forest (Trivedi, 2020):

Given: nT- number of training examples, xi-number of all features; Xe -number of features selected for Ensembles; mi - number of all Ensemble members. The Random Forest is constructed using mi trees.

For mi classifiers, the below steps are performed:

- **First:** Bagging method is performed to create nT number of samples and each sample is used as training data for decision dump classifier.
- **Second:** For constructing trees for random forest, random features are selected using Gini index and the tree grows without trimming.
- **Third:** Training samples from bagging methods are applied on decision trees mi to generate trained models. Further, trained models are evaluated using voting mechanism and a final classification decision is taken.

In a regression scenario, the final prediction is obtained by averaging the outputs of individual decision trees. By aggregating predictions from multiple trees, Random Forest effectively reduces variance in prediction errors and outperforms standalone decision trees. This bagging technique leverages the diversity of the ensemble to enhance predictive accuracy

Random Forest is considered as one of the best algorithms currently which is not sensitive to multicollinearity, and the results are relatively robust to missing data and unbalanced data. In addition, it can well predict the role of up to thousands of explanatory variables.

4.3.2 BOOSTING

The basic notion underlying boosting is that we may "boost," as the name implies, a weak algorithm into a stronger algorithm with a far more potent prediction ability. Currently, the algorithm performs just marginally better than random chance.

Unlike Bagging, Boosting creates different base learners by sequentially reweighting the instances in the training dataset. Each instance misclassified by the previous base learner will get a larger weight in the next round of training.

The first step in the boosting process is choosing a base learning algorithm, often known as a weak learner. Weak learners are models, such as shallow-depth decision trees, linear models, or even basic rule-based models, that are relatively simple and have little predictive ability on their own. To ensure that they can be trained rapidly and are less prone to overfit the training data, these weak learners are selected.

Every boosting iteration involves assigning a weight to each training instance according to its significance. At the beginning, each training instance is assigned the same weight. The weights of cases that are incorrectly classified increase while the weights of successfully classified instances drop as boosting advances. By making sure that incoming poor learners focus more on the cases that are challenging to categorize, this adaptive weighting strategy enhances the ensemble's performance as a whole.

Boosting aggregates the predictions of several weak learners trained on different instances to get a final prediction. Weak learners usually combine their predictions using a weighted total, where each weak learner's performance during training determines the weights. In the final ensemble, weak learners who do well on the training data are assigned larger weights, while those who do poorly are either assigned lower weights or might be eliminated entirely.

Input: Sample distribution $\mathcal{D}$;
Base learning algorithm $\mathfrak{L}$;
Number of learning rounds T.

Process:

1. $\mathcal{D}_1 = \mathcal{D}$. % Initialize distribution
2. **for** $t = 1, \ldots, T$:
3. $h_t = \mathfrak{L}(\mathcal{D}_t)$; % Train a weak learner from distribution $\mathcal{D}_t$
4. $\epsilon_t = P_{\boldsymbol{x} \sim D_t}(h_t(\boldsymbol{x}) \neq f(\boldsymbol{x}))$; % Evaluate the error of h_t
5. $\mathcal{D}_{t+1} = Adjust\_Distribution(\mathcal{D}_t, \epsilon_t)$
6. **end**

Output: $H(\boldsymbol{x}) = Combine\_Outputs(\{h_1(\boldsymbol{x}), \ldots, h_t(\boldsymbol{x})\})$

Figure 8 *Boosting Algurithm*

Source: Author Preparation

4.3.2.1 GRADIENT BOOSTING AND XGBOOST

To understand XGBoost, we need first to introduce Gradient Boosting Tree. Gradient Boosting Tree is a boosting ensemble technique utilized for both regression and classification problems. It uses fixed-size decision trees to represent weak learners. In order to create a strong predictive model, Gradient Boosting builds a collection of weak prediction models—in this case, decision trees—that make very few assumptions about the data. GBT forms a powerful predictive model because, in contrast to certain other machine learning algorithms, it makes very few assumptions about the input.

The fundamental idea behind Gradient Boosting is to gradually develop trees that repair mistakes made by earlier trees in order to improve performance iteratively. In order to minimize a differentiable loss function, model parameters are adjusted iteratively. This loss function measures the difference between the model's predicted values and the dataset's observed values. GBT improves both the model's fit to the training set and its generalization to new, unobserved data by iteratively minimizing this loss function.

Given a dataset $D = \{x_i, y_i\}_1^N$, the goal of gradient boosting is to find an approximation, $\hat{F}(x)$, of the function $F^*(x)$, which maps instances to their output values y, by minimizing the expected value of a given loss function, $L(y, F(x))$. Gradient boosting builds an additive approximation of $F^*(x)$ as a weighted sum of functions $F_m(x) = F_{m-1}(x) + \rho_m h_m(x)$, where ρ_m is the weight of the m^{th} function, $h_m(x)$. These functions are the models of the ensemble technique. Subsequent models are expected to minimize $\big(\rho_m, h_m(x)\big) = arg\min_{\rho,h} \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + \rho h(x_i))$.

Extreme Gradient Boosting (XGBoost) is an optimized continuation of Gradient Boosting, aiming to address a significant limitation of Gradient Boosting: the exhaustive consideration of potential splits at each node for branch creation (Chen & Guestrin, 2016). This exhaustive approach often leads to increased computational complexity and potential overfitting. XGBoost (Qiu, 2019) tackles this issue by introducing regularization terms that control the complexity of individual trees, thereby improving the model's generalization capabilities.

The objective function is:

$$L(\emptyset) = \sum_{i=1}^{I} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k) \qquad (4)$$

$\sum_{i=1}^{I} l(\hat{y}_i, y_i)$ is the sum of differentiable convex loss function of all basic learners that measure the difference between the prediction $\hat{y}_i$ and the target y_i. The second term Ω penalizes the complexity of the model (i.e., the regression tree function). $\sum_{k=1}^{K} \Omega(f_k)$ is a sum of all basic learners' regular functions.

$$\Omega(f) = \Upsilon T + \frac{1}{2}\lambda \|w\|^2 = \Upsilon T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \qquad (5)$$

where:

- γ: Complexity of each leaf;
- T: Number of leaves in the tree;
- λ: Parameter to scale the penalty;
- w: Vector of scores on leaves to level the penalty.

After deducing the above formulas, we have a detailed understanding of the t-th lifting process of XGBoost.

First, we initialize a basic learner $\hat{y}_i^{(0)} = 0$, which means that after the 0th round of promotion, the output value of the 0th base learner for each sample is 0. Therefore, in the first round of promotion, the residuals produced by the real values subtracting the prediction values from the model are the real values of the sample itself. Therefore, the first round of lifting will be based on the values of the samples as a data set. When the first round of lifting is completed, we can get the output values of the model for each sample by calculating. The output values are obtained by summing up the output values given by the two basic learners (the 0th and the 1st learners) after the first lifting. Each subsequent iteration is a repetition of the above process.

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t-1} f_k(x_i) + f_t(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (6)$$

f_t is the t – th tree structure. Every time a tree is added to the model, its loss function will change. The t −1 trees will be added into the model after finishing training.

With all this information, The predicting function of XGBoost is:

$$\hat{y}_i = \emptyset(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F} \quad (7)$$

where:

- $f_k(x_i)$ is the output value for the sample x_i from the k – th basic predictor;
- $\hat{y}_i$ is the sum of output values given by K basic learners to sample x_i;
- $\mathcal{F} = \{f(x) = w_q(x)\}(q: \mathbb{R}^m \to T, w \in \mathbb{R}^T)$.

5. PERFORMANCE MEASURES

In this thesis we will evaluate the models by two different types of performance measures (Fawcett, 2004):

- Discriminatory ability;
- Accuracy;
- Performance.

To determine these measures, we will delve into the foundational aspects of classification problems involving only two classes. Here, every instance is assigned to either a positive (p) or negative (n) class label. Imagine it categorizes applicants into approved or denied. A classification model, commonly referred to as a classifier, serves as a mechanism to predict the class to which each instance belongs. To distinguish between the actual class and the predicted class, we adopt labels such as “P” for positive and “N” for negative predictions generated by the model. This enables us to differentiate between the true nature of an instance and the classification assigned by the model.

Upon evaluation of a classifier using a given entity, four distinct outcomes may arise. If the entity is genuinely positive and is correctly classified as positive, it contributes to a count of true positives. Conversely, if it is mistakenly classified as negative, it contributes to a count of false negatives. Similarly, if the entity is genuinely negative and is accurately classified as negative, it adds to the count of true negatives; however, if it is erroneously classified as positive, it contributes to a count of false positives.

When extending this evaluation to encompass a classifier and a set of entities, commonly referred to as the test set, a two-by-two confusion matrix, also known as a contingency table, can be constructed. This matrix serves as a concise representation of the disposition of the entities within the set concerning their true and predicted classes. Importantly, the confusion matrix lays the groundwork for various sophisticated evaluation metrics widely employed in the domain of credit scoring models.

Table 2 *Confusion Matrix*

Source: Authors Preparation

	Predicted Positive (P)	Predicted Negative (N)
Actually Positive (p)	True Positive	False Negative
Actually Negative (n)	False Positive	True Negative

Table 2 shows a confusion matrix and we can compute several common metrics that can be calculated from it. The numbers along the major diagonal represent the correct decisions made, and the numbers off this diagonal represent the errors—the confusion—between the various classes.

The above matrix can be interpreted as follows:

- True Negative - Predicted as non-default and it is actually non-default;
- False Positive - Predicted as default when it is actually non default;
- False Negative - Predicted as non-default when it is actually default;
- True positive - Predicted as default and it is actually default.

True Positive Rate (Recall)

This metric measures how many positive instances are correctly predicted amongst all positive samples. In other words, recall answers the question of what proportion of positive predictions (default) were actually correct.

$$True\ Positive\ Rate = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (8)$$

True Negative Rate (Specificity)

The percentage of all negative events that are classified as negative is measured using this evaluation metric. Stated differently, specificity is the percentage of non-defaults, or negative forecasts, that turned out to be true.

$$True\ Negative\ Rate = \frac{True\ Negatives}{False\ positives + True\ Negatives} \quad (9)$$

False Positive Rate (Type I Error)

Identifies the good candidates who are expected to be bad. This mistake can be computed using the following formula, or as 1-Specificity:

$$False\ Positive\ Rate = \frac{False\ Positive}{False\ Positive+True\ Negative} \quad (10)$$

False Negative Rate (Type II Error)

Defines the bad applicants that are predicted to be good. In the presence of a type II error, a default applicant is misclassified as non-default. It can be calculated as 1- Recall or by the formula given as follows:

$$False\ Negative\ Rate = \frac{False\ Negative}{True\ Positive+False\ Negaive} \quad (11)$$

Accuracy

The accuracy metric calculates the proportion of positively and negatively classed events that were accurately classified out of all the events that were examined. One way to calculate accuracy is as follows:

$$Accuracy = \frac{True\ Positive+True\ Negative}{True\ Positive+True\ Negative+False\ Positive+False\ Negative} \quad (12)$$

ROC curve and AUC

Receiver Operating Characteristics (ROC) graphs are a useful technique for organizing classifiers and visualizing model´s performance. In recent years, ROC graphs have been increasingly adopted in the machine learning and data mining research communities.
One of the earliest adopters of ROC graphs in machine learning was (Spackman, 1989) who demonstrated the value of ROC curves in evaluating and comparing algorithms. Recent years have seen an increase in the use of ROC graphs in the machine learning community.
ROC graphs are two-dimensional graphs in which true positive rate is plotted on the Y axis and false positive rate is plotted on the X axis.
The optimal classifier would produce a data point in the upper left corner of the ROC space, signifying 100% sensitivity and 100% specificity (i.e., no false negatives or positives). An entirely arbitrary guess would be represented by a point along the diagonal. Since the diagonal splits the ROC space, data points above the diagonal line indicate good classification results (better than random chance), and those below it indicate poor results (worse than random).
An ROC curve is a two-dimensional depiction of classifier performance. To compare classifiers we may want to reduce ROC performance to a single scalar value representing expected performance. A common method is to calculate the area under the ROC curve, abbreviated

AUC. AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance.

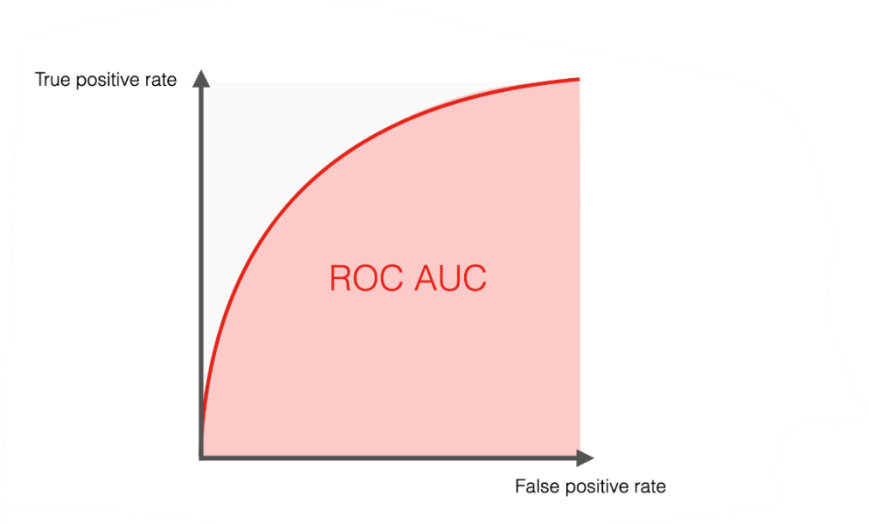


Figure 9 *ROC Curve*

Source: EvidentlyAI

The classifier is perfectly competent to distinguish between the positive and negative class points when AUC = 1. All positives would be projected as negatives and all negatives as positives if the AUC was 0, which would mean that the classifier would be incorrect in all of its predictions. When the AUC value is 0.5, it means that the ROC curve will match the diagonal, indicating that the classifier is not selective and will not outperform random choice. There is a good probability that the classifier will be able to distinguish between the positive and negative classes when the AUC falls between 0.5 and 1. This happens because the classifier is able to predict a higher number of true negatives and true positives than false negatives and false positives.

6. DATA PROPERTIES AND DATA PREPOCESSING

In this section we present the data set used in our paper, which is publicly provided by Freddie Mac (Mac, 2019) as the complete exploratory analysis of the data and the pre-processing steps using Python. The dataset covers approximately 250000 fixed-rate mortgages originated between January 2018 and Dezember 2022.

6.1 DATA DICTIONARY

The dataset provided has two sets of data, Loan-level origination files, and monthly loan performance:

- Origination Data file

The origination Data file contains loan-level origination information for all the loans.

Table 3 *Original Data Features*

Source: Fredie Mac

Column Position	Formal Name and Definition	Valid Values	Type
1	**CREDIT SCORE** - A number, prepared by third parties, summarizing the borrower's creditworthiness, which may be indicative of the likelihood that the borrower will timely repay future obligations. Generally, the credit score disclosed is the score known at the time of acquisition and is the score used to originate the mortgage.	• 300 - 850 • 9999 = Not Available, if Credit Score is < 300 or > 850.	Numeric
2	**FIRST PAYMENT DATE** - The date of the first scheduled mortgage payment due under the terms of the mortgage note.	YYYYMM	Date
3	**FIRST TIME HOMEBUYER FLAG** - Indicates whether the Borrower, or one of a group of Borrowers, is an individual who (1) is purchasing the mortgaged property, (2) will reside in the mortgaged property as a primary residence and (3) had no ownership interest (sole or joint) in a residential property during the three-year period preceding the date of the purchase of the mortgaged property. With certain limited exceptions, a displaced homemaker or single parent may also be considered a First-Time Homebuyer if the individual had no ownership interest in a residential property during the preceding three-year period other than an ownership interest in the marital residence with a spouse.	• Y = Yes • N = No • 9 = Not Available or Not Applicable	Alpha
4	**MATURITY DATE** - The month in which the final monthly payment on the mortgage is scheduled to be made as stated on the original mortgage note.	• YYYYMM	Date
5	**METROPOLITAN STATISTICAL AREA (MSA) OR METROPOLITAN DIVISION** - This disclosure will be based on the designation of the Metropolitan Statistical Area or Metropolitan Division as of the date of issuance. Metropolitan Statistical Areas (MSAs) are defined by the United States Office of Management and Budget (OMB) and have at least one urbanized area with a population of 50,000 or more inhabitants. An MSA containing a single core with a population of 2.5 million or more may be divided into smaller groups of counties that OMB refers to as Metropolitan Divisions. If an MSA applies to a mortgaged property, the applicable five-digit value is disclosed; however, if the mortgaged	• Metropolitan Division or MSA Code. • Space (5) = Indicates that the area in which the mortgaged property is located is a) neither an MSA nor a Metropolitan Division, or b) unknown.	Numeric

	property also falls within a Metropolitan Division classification, the applicable five-digit value for the Metropolitan Division takes precedence and is disclosed instead. This disclosure will not be updated to reflect any subsequent changes in designations of MSAs, Metropolitan Divisions or other classifications. Null indicates that the area in which the mortgaged property is located is (a) neither an MSA nor a Metropolitan Division, or (b) unknown.		
6	**MORTGAGE INSURANCE PERCENTAGE (MI %)** - The percentage of loss coverage on the loan, at the time of Freddie Mac's purchase of the mortgage loan that a mortgage insurer is providing to cover losses incurred as a result of a default on the loan. Only primary mortgage insurance that is purchased by the Borrower, lender or Freddie Mac is disclosed. Mortgage insurance that constitutes "credit enhancement" that is not required by Freddie Mac's Charter is not disclosed. Amounts of mortgage insurance reported by Sellers that are less than 1% or greater than 55% will be disclosed as "Not Available," which will be indicated 999. No MI will be indicated by three zeros.	• 1% - 55% • 000 = No MI • 999 = Not Available	Numeric
7	**NUMBER OF UNITS** - Denotes whether the mortgage is a one-, two-, three-, or four-unit property.	• 1 = one-unit • 2 = two-unit • 3 = three-unit • 4 = four-unit • 99 = Not Available	Numeric
8	**OCCUPANCY STATUS** - Denotes whether the mortgage type is owner occupied, second home, or investment property.	• P = Primary Residence • I = Investment Property • S = Second Home • 9 = Not Available	Alpha
9	**ORIGINAL COMBINED LOAN-TO-VALUE (CLTV)** – In the case of a purchase mortgage loan, the ratio is obtained by dividing the original mortgage loan amount on the note date plus any secondary mortgage loan amount disclosed by the Seller by the lesser of the mortgaged property's appraised value on the note date or its purchase price. In the case of a refinance mortgage loan, the ratio is obtained by dividing the original mortgage loan amount on the note date plus any secondary mortgage loan amount disclosed by the Seller by the mortgaged property's appraised value on the note date. If the secondary financing amount disclosed by the Seller includes a home equity line of credit, then the CLTV calculation reflects the disbursed amount at closing of the first lien mortgage loan, not the maximum loan amount available under the home equity line of credit. In the case of a seasoned mortgage loan, if the Seller cannot warrant that the value of the mortgaged property has not declined since the note date, Freddie Mac requires that the Seller must provide a new appraisal value, which is used in the CLTV calculation. In certain cases, where the Seller delivered a loan to Freddie Mac with a special code indicating additional secondary mortgage loan amounts, those amounts may have been included in the CLTV calculation. If the CLTV is < LTV, set the CLTV to 'Not Available.' This disclosure is subject to the widely varying standards originators use to verify Borrowers' secondary mortgage loan amounts and will not be updated.	2018Q1 and prior: • 6% - 200% • 999 = Not Available 2018Q2 and later: • 1% - 998% • 999 = Not Available HARP ranges: • 1% - 998% • 999 = Not Available	Numeric
10	**ORIGINAL DEBT-TO-INCOME (DTI) RATIO** - Disclosure of the debt to income ratio is based on (1) the sum of the borrower's monthly debt payments, including monthly housing expenses that incorporate the mortgage payment the borrower is making at the time of the delivery of the mortgage loan to Freddie Mac, divided by (2) the total monthly income used to underwrite the loan as of the date of the origination of the such loan. Ratios greater than 65% are indicated that data is Not Available. All loans in the HARP dataset will be disclosed as Not Available. This disclosure is subject to the widely varying standards originators use to verify Borrowers' assets and liabilities and will not be updated.	•0%<DTI<=65% • 999 = Not Available HARP ranges: • 999 = Not Available	Numeric

11	**ORIGINAL UPB** - The UPB of the mortgage on the note date.	• Amount will be rounded to the nearest $1,000.	Numeric
12	**ORIGINAL LOAN-TO-VALUE (LTV)** - In the case of a purchase mortgage loan, the ratio obtained by dividing the original mortgage loan amount on the note date by the lesser of the mortgaged property's appraised value on the note date or its purchase price. In the case of a refinance mortgage loan, the ratio obtained by dividing the original mortgage loan amount on the note date and the mortgaged property's appraised value on the note date. In the case of a seasoned mortgage loan, if the Seller cannot warrant that the value of the mortgaged property has not declined since the note date, Freddie Mac requires that the Seller must provide a new appraisal value, which is used in the LTV calculation. For loans in the non HARP dataset, ratios below 6% or greater than 105% will be disclosed as "Not Available," indicated by 999. For loans in the HARP dataset, LTV ratios greater than 999% will be disclosed as Not Available.	2018Q1 and prior: • 6% - 105% • 999 = Not Available 2018Q2 and later: • 1% - 998% • 999 = Not Available HARP ranges: • 1% - 998% • 999 = Not Available	Numeric
13	**ORIGINAL INTEREST RATE** - The original note rate as indicated on the mortgage note.		Numeric Literal decimal
14	**CHANNEL** - Disclosure indicates whether a Broker or Correspondent, as those terms are defined below, originated or was involved in the origination of the mortgage loan. If a Third Party Origination is applicable, but the Seller does not specify Broker or Correspondent, the disclosure will indicate "TPO Not Specified". Similarly, if neither Third Party Origination nor Retail designations are available, the disclosure will indicate "TPO Not Specified." If a Broker, Correspondent or Third Party Origination disclosure is not applicable, the mortgage loan will be designated as Retail, as defined below. Broker is a person or entity that specializes in loan originations, receiving a commission (from a Correspondent or other lender) to match Borrowers and lenders. The Broker performs some or most of the loan processing functions, such as taking loan applications, or ordering credit reports, appraisals and title reports. Typically, the Broker does not underwrite or service the mortgage loan and generally does not use its own funds for closing; however, if the Broker funded a mortgage loan on a lender's behalf, such a mortgage loan is considered a "Broker" third party origination mortgage loan. The mortgage loan is generally closed in the name of the lender who commissioned the Broker's services. Correspondent is an entity that typically sells the Mortgages it originates to other lenders, which are not Affiliates of that entity, under a specific commitment or as part of an ongoing relationship. The Correspondent performs some, or all, of the loan processing functions, such as: taking the loan application; ordering credit reports, appraisals, and title reports; and verifying the Borrower's income and employment. The Correspondent may or may not have delegated underwriting and typically funds the mortgage loans at settlement. The mortgage loan is closed in the Correspondent's name and the Correspondent may or may not service the mortgage loan. The Correspondent may use a Broker to perform some of the processing functions or even to fund the loan on its behalf; under such circumstances, the mortgage loan is considered a "Broker" third party origination mortgage loan, rather than a "Correspondent" third party origination mortgage loan. Retail Mortgage is a mortgage loan that is originated, underwritten and funded by a lender or its Affiliates. The mortgage loan is closed in the name of the lender or its Affiliate and if it is sold to Freddie Mac, it is sold by the lender or its Affiliate that originated it. A mortgage loan that a Broker or Correspondent completely or partially originated, processed, underwrote, packaged, funded or closed is not considered a Retail mortgage loan. For purposes of the definitions of Correspondent and Retail, "Affiliate" means any entity that is related to another party as a consequence of the entity, directly or indirectly, controlling the other party, being controlled by the other party, or being under common control with the other party.	• R = Retail • B = Broker • C = Correspondent • T = TPO Not Specified • 9 = Not Available	Alpha
15	**PREPAYMENT PENALTY MORTGAGE (PPM) FLAG** - Denotes whether the mortgage is a PPM. A PPM is a mortgage with respect	• Y = PPM	Alpha

	to which the borrower is, or at any time has been, obligated to pay a penalty in the event of certain repayments of principal.	• N = Not PPM	
16	**AMORTIZATION TYPE** - Denotes that the product is a fixed-rate mortgage or adjustable-rate mortgage.	• FRM – Fixed Rate Mortgage • ARM – Adjustable Rate Mortgage	Alpha
17	**PROPERTY STATE** - A two-letter abbreviation indicating the state or territory within which the property securing the mortgage is located.	• AL, TX, VA, etc.	Alpha
18	**PROPERTY TYPE** - Denotes whether the property type secured by the mortgage is a condominium, leasehold, planned unit development (PUD), cooperative share, manufactured home, or Single-Family home. If the Property Type is Not Available, this will be indicated by 99.	• CO = Condo • PU = PUD • MH = Manufactured Housing • SF = Single-Family • CP = Co-op • 99 = Not Available	Alpha
19	**POSTAL CODE** – The postal code for the location of the mortgaged property	• ###00, where "###" represents the first three digits of the 5-digit postal code • Space (5) = Unknown	Numeric
20	**LOAN SEQUENCE NUMBER** - Unique identifier assigned to each loan.	PYYQnXXXXXXX • Product F = FRM and A = ARM; • YYQn = origination year and quarter; and, • XXXXXXX = randomly assigned digits	Alpha-numeric
21	**LOAN PURPOSE** - Indicates whether the mortgage loan is a Cash-out Refinance mortgage, No Cash-out Refinance mortgage, or a Purchase mortgage. Generally, a Cash-out Refinance mortgage loan is a mortgage loan in which the use of the loan amount is not limited to specific purposes. A mortgage loan placed on a property previously owned free and clear by the Borrower is always considered a Cash-out Refinance mortgage loan. Generally, a No Cash-out Refinance mortgage loan is a mortgage loan in which the loan amount is limited to the following uses: Pay off the first mortgage, regardless of its age Pay off any junior liens secured by the mortgaged property, that were used in their entirety to acquire the subject property Pay related closing costs, financing costs and prepaid items, and Disburse cash out to the Borrower (or any other payee) not to exceed 2% of the new refinance mortgage loan or $2,000, whichever is less. As an exception to the above, for construction conversion mortgage loans and renovation mortgage loans, the amount of the interim construction financing secured by the mortgaged property is considered an amount used to pay off the first mortgage. Paying off unsecured liens or construction costs paid by the Borrower outside of the secured interim construction financing is considered cash out to the Borrower, if greater than $2000 or 2% of loan amount. This disclosure is subject to various special exceptions used by Sellers to determine whether a mortgage loan is a No Cash-out Refinance mortgage loan.	• P = Purchase • C = Refinance - Cash Out • N = Refinance - No Cash Out • R = Refinance - Not Specified • 9 =Not Available	Alpha
22	**ORIGINAL LOAN TERM** - A calculation of the number of scheduled monthly payments of the mortgage based on the First Payment Date and Maturity Date.	• Calculation: (Loan Maturity Date (MM/YY) – Loan First Payment Date (MM/YY) + 1)	Numeric

23	**NUMBER OF BORROWERS** - The number of Borrower(s) who are obligated to repay the mortgage note secured by the mortgaged property. Disclosure denotes only whether there is one borrower, or more than one borrower associated with the mortgage note. This disclosure will not be updated to reflect any subsequent assumption of the mortgage note.	2018Q1 and prior: • 01 = 1 borrower • 02 = > 1 borrowers • 99 = Not Available 2018Q2 and later: • 01 = 1 borrower • 02 = 2 borrowers • 03 = 3 borrowers … • 09 = 9 borrowers • 10 = 10 borrowers • 99 = Not Available	Numeric
24	**SELLER NAME** - The entity acting in its capacity as a seller of mortgages to Freddie Mac at the time of acquisition. Seller Name will be disclosed for sellers with a total Original UPB representing 1% or more of the total Original UPB of all loans in the Dataset for a given calendar quarter. Otherwise, the Seller Name will be set to "Other Sellers".	Name of the seller, or "Other Sellers"	Alpha-numeric
25	**SERVICER NAME** - The entity acting in its capacity as the servicer of mortgages to Freddie Mac as of the last period for which loan activity is reported in the Dataset. Servicer Name will be disclosed for servicers with a total Original UPB representing 1% or more of the total Original UPB of all loans in the Dataset for a given calendar quarter. Otherwise, the Servicer Name will be set to "Other Servicers".	Name of the servicer, or "Other Servicers"	Alpha-numeric
26	**SUPER CONFORMING FLAG** – For mortgages that exceed conforming loan limits with origination dates on or after 10/1/2008 and were delivered to Freddie Mac on or after 1/1/2009	• Y = Yes • Space (1) = Not Super Conforming	Alpha
27	**PRE-RELIEF REFINANCE LOAN SEQUENCE NUMBER –** The Loan Sequence Number link that associates a Relief Refinance loan to the Loan Sequence Number assigned to the loan from which it was refinanced within in the Single-Family Loan Level Dataset. Note: Populated only for loans where the Relief Refinance Indicator is set to Y. All other loans will be blank.	PYYQnXXXXXXX • Product F = FRM and A = ARM; • YYQn = origination year and quarter; and, • XXXXXXX = randomly assigned digits	Alpha-numeric
28	**PROGRAM INDICATOR –** The indicator that identifies if a loan participates in and of the Freddie Mac programs listed in the valid values. Note: The standard dataset discloses these enumerations for loans originated on or after March 1, 2015. The Non-Standard dataset discloses enumerations for loans originated under the HP program between 1999 and February 28, 2015. Underwriting standards for Home Possible prior to March 1, 2015 may be different than the current standards.	H = Home Possible F = HFA Advantage R= Refi Possible 9 = Not Available or Not Applicable	Alpha-numeric
29	**RELIEF REFINANCE INDICATOR –** Indicator that identifies whether the loan is part of Freddie Mac's Relief Refinance Program. Loans which are both a Relief Refinance and have an Original Loan-to-Value above 80 are HARP loans.	Y = Relief Refinance Loan Space = Non-Relief Refinance loan	Alpha
30	**PROPERTY VALUATION METHOD –** The indicator denoting which method was used to obtain a property appraisal, if any. Note: Populated for loans originated on or after 1/1/2017.	1 = ACE Loans 2 = Full Appraisal 3 = Other Appraisals (Desktop, driveby, external, AVM) 4 = ACE + PDR 9 = Not Available	Numeric
31	**INTEREST ONLY INDICATOR (I/O INDICATOR) -** The indicator denoting whether the loan only requires interest payments for a specified period beginning with the first payment date.	Y = Yes N = No	Alpha
32	**MI CANCELLATION INDICATOR:** The indicator denoting if the mortgage insurance has been reported as cancelled after the time of Freddie Mac's purchase of the mortgage loan. If a loan did not have mortgage insurance at the time of Freddie Mac's purchase of the mortgage loan, then this field will be disclosed as "Not Applicable." The Mortgage Insurance Cancellation Indicator will remain constant beginning in the month in which the loan is removed from the Reference Pool.	Y = Yes, MI has been cancelled N = No, MI has not been cancelled 7 = Not Applicable 9 = Not Disclosed	Alpha Numeric

- Monthly Performance Data File

The monthly performance data file contains monthly loan-level credit performance and actual loss data for each loan, starting from the time of loan acquisition by Freddie Mac until the earlier of a termination event or the Performance Cutoff Date, which is the last period of performance data available for any loan in the Dataset.

Table 4 *Monthly Performance Features*

Source: FredieMac

Column Position	Formal Name and Definition	Valid Values	Type
1	**LOAN SEQUENCE NUMBER** - Unique identifier assigned to each loan.	PYYQnXXXXXXX • Product F = FRM and A = ARM; • YYQn = origination year and quarter; and, • XXXXXXX = randomly assigned digits	Alpha-numeric
2	**MONTHLY REPORTING PERIOD** – The as-of month for loan information contained in the loan record.	YYYYMM	Date
3	**CURRENT ACTUAL UPB** - The Current Actual UPB reflects the mortgage ending balance as reported by the servicer for the corresponding monthly reporting period. For fixed rate mortgages, this UPB is derived from the mortgage balance as reported by the servicer and includes any scheduled and unscheduled principal reductions applied to the mortgage. For mortgages with loan modifications, as indicated by "Y" in the Modification Flag field, the current actual unpaid principal balance may or may not include partial principal forbearance. If applicable, for loans with partial principal forbearance, the current actual unpaid principal balance equals the sum of interest bearing UPB (the amortizing principal balance of the mortgage) and the deferred UPB (the principal forbearance balance). Current UPB will be rounded to the nearest $1,000 for the first 6 months after origination date. This was previously reported as zero for the first 6 months after the origination date.	**Calculation:** (interest bearing UPB) + (non-interest bearing UPB)	Numeric Literal decimal
4	**CURRENT LOAN DELINQUENCY STATUS** – A value corresponding to the number of days the borrower is delinquent, based on the due date of last paid installment ("DDLPI") reported by servicers to Freddie Mac, and is calculated under the Mortgage Bankers Association (MBA) method. If a loan has been acquired by REO, then the Current Loan Delinquency Status will reflect the value corresponding to that status (instead of the value corresponding to the number of days the borrower is delinquent).	• XX = Unknown • 0 = Current, or less than 30 days past due • 1 = 30-59 days delinquent • 2 = 60 – 89 days delinquent • 3 = 90 – 119 days delinquent • And so on… • RA = REO Acquisition • Space (3) = Unavailable	Alpha-numeric
5	**LOAN AGE** - The number of scheduled payments from the time the loan was originated or modified up to and including the performance cutoff date.	**Calculation – Non-modified Loans:** ((Monthly Reporting Period) – Loan First Payment Date (MM/YY)) +1 month **Calculation – Modified Loans:** ((Monthly Reporting Period) - Modification First Payment Date (MM/YY)) +1 month	Numeric
6	**REMAINING MONTHS TO LEGAL MATURITY** - The remaining number of months to the mortgage maturity date.	**Calculation:** (Maturity Date (MM/YY) – Monthly Reporting Period (MM/YY)	Numeric

	For mortgages with loan modifications, as indicated by "Y" in the Modification Flag field, the calculation uses the modified maturity date.		
7	**DEFECT SETTLEMENT DATE:** For mortgages that experienced a credit event and for which Freddie Mac has: (I) previously determined the existence of an Unconfirmed Underwriting defect, the date on which there is the occurrence of any of the following: (x) such mortgage is repurchased by the related seller or servicer, (y) in lieu of repurchase, an alternative remedy (such as indemnification) is mutually agreed upon by both Freddie Mac and the seller or servicer or (z) Freddie Mac in its sole discretion elects to waive the enforcement of a remedy against the seller or servicer in respect of such Unconfirmed Underwriting Defect; or; (II) previously determined the existence of an Unconfirmed Servicing Defect and the existence of a Major Servicing Defect, the date on which there is the occurrence of any of the following: (x) the related servicer repurchased such mortgage or made Freddie Mac whole resulting in a full recovery of losses incurred ("Make Whole") or (y) the party responsible for the representations and warranties and/or servicing obligations or liabilities with respect to the mortgage becomes subject to a bankruptcy, an insolvency proceeding or a receivership. Loans covered under any Origination Rep and Warranty Settlements Servicing Settlements are also included in this category.	YYYYMM	Date
8	**MODIFICATION FLAG** – For mortgages with loan modifications, indicates that the loan has been modified.	• Y = Current Period Modification • P = Prior Period Modification • Space (1) = Not Modified	Alpha
9	**ZERO BALANCE CODE** - A code indicating the reason the loan's balance was reduced to zero.	• 01 = Prepaid or Matured (Voluntary Payoff) • 02 = Third Party Sale • 03 = Short Sale or Charge Off • 96 = Repurchase prior to Property Disposition • 09 = REO Disposition • 15 = Whole Loan sales • 16 = Reperforming sales securitizaitons	Numeric
10	**ZERO BALANCE EFFECTIVE DATE** - The date on which the event triggering the Zero Balance Code took place.	YYYYMM Space (6) = Not Applicable	Date
11	**CURRENT INTEREST RATE** - Reflects the current interest rate on the mortgage note, taking into account any loan modifications.		Numeric Literal Decimal
12	**CURRENT DEFERRED UPB**: The current non-interest bearing UPB of the modified mortgage.	$ Amount. Non-Interest Bearing UPB.	Numeric
13	**DUE DATE OF LAST PAID INSTALLMENT (DDLPI)**: The due date that the loan's scheduled principal and interest is paid through, regardless of when the installment payment was actually made.	YYYYMM	Date
14	**MI RECOVERIES** - Mortgage Insurance Recoveries are proceeds received by Freddie Mac in the event of credit losses. These proceeds are based on claims under a mortgage insurance policy. Note: the MI Recoveries field will be set to zero for loans with a Defect Settlement Date value populated.	$ Amount. MI Recoveries.	Numeric Literal Decimal
15	**NET SALE PROCEEDS** - The amount remitted to Freddie Mac resulting from a property disposition or loan sale (which in the case of bulk sales, may be an allocated amount) once allowable selling expenses have been deducted from the gross sales proceeds. A value of "U" indicates that the amount is unknown. Note: the Net Sale Proceeds field will be set to zero for loans with a Defect Settlement Date value populated.	$ Amount. Gross Sale Proceeds – Allowable Selling Expenses. U = Unknown	Alpha-numeric Literal Decimal

16	**NON MI RECOVERIES**: Non-MI Recoveries are proceeds received by Freddie Mac based on repurchase/make whole proceeds, non-sale income such as refunds (tax or insurance), hazard insurance proceeds, rental receipts, positive escrow and/or other miscellaneous credits. Note: the Non MI Recoveries field will be set to zero for loans with a Defect Settlement Date value populated.	$ Amount. Non MI Recoveries.	Numeric Literal Decimal
17	**EXPENSES** - Expenses will include allowable expenses that Freddie Mac bears in the process of acquiring, maintaining and/ or disposing a property (excluding selling expenses, which are subtracted from gross sales proceeds to derive net sales proceeds). This is an aggregation of Legal Costs, Maintenance and Preservation Costs, Taxes and Insurance, and Miscellaneous Expenses. Note: the Expenses field will be set to zero for loans with a Defect Settlement Date value populated.	$ Amount. Allowable Expenses.	Numeric Literal Decimal
18	**LEGAL COSTS -** The amount of legal costs associated with the sale of a property (but not included in Net Sale Proceeds). Prior to population of a Zero Balance Code equal to 02, 03, 09, or 15, this field will be populated as "Not Applicable," Following population of a Zero Balance Code equal to 02, 03, 09, or 15, this field will be updated (as applicable) to reflect the cumulative total. Space (12) – Not applicable Note: the Legal Costs field will be set to zero for loans with a Defect Settlement Date value populated.	$ Amount	Numeric Literal Decimal
19	**MAINTENANCE AND PRESERVATION COSTS –**The amount of maintenance, preservation, and repair costs, including but not limited to property inspection, homeowner's association, utilities, and REO management, that is associated with the sale of a property (but not included in Net Sale Proceeds). Prior to population of a Zero Balance Code equal to 02, 03, 09, or 15, this field will be populated as "Not Applicable," Following population of a Zero Balance Code equal to 02, 03, 09, or 15, this field will be updated (as applicable) to reflect the cumulative total. Space (12) – Not applicable Note: the Maintenance and Preservation Costs field will be set to zero for loans with a Defect Settlement Date value populated.	$ Amount	Numeric Literal Decimal
20	**TAXES AND INSURANCE** – The amount of taxes and insurance owed that are associated with the sale of a property (but not included in Net Sale Proceeds). Prior to population of a Zero Balance Code equal to 02, 03, 09, or 15, this field will be populated as "Not Applicable,". Following population of a Zero Balance Code equal to 02, 03, 09, or 15, this field will be updated (as applicable) to reflect the cumulative total. Space (12) – Not applicable Note: the Taxes and Insurance field will be set to zero for loans with a Defect Settlement Date value populated.	$ Amount	Numeric Literal Decimal
21	**MISCELLANEOUS EXPENSES** - Miscellaneous expenses associated with the sale of a property (but not included in Net Sale Proceeds). Prior to population of a Zero Balance Code equal to 02, 03, 09, or 15, this field will be populated as "Not Applicable,". Following population of a Zero Balance Code equal to 02, 03, 09, or 15, this field will be updated (as applicable) to reflect the cumulative total. Space (12) – Not applicable Note: the Miscellaneous Expenses field will be set to zero for loans with a Defect Settlement Date value populated.	$ Amount	Numeric Literal Decimal
22	**ACTUAL LOSS CALCULATION** Actual Loss is calculated using the below approach: Actual Loss = (Default UPB – Net Sale_Proceeds) + Delinquent Accrued Interest - Expenses – MI Recoveries – Non MI Recoveries. Delinquent Accrued Interest = (Default_Upb – Non Interest bearing UPB) * Min(Current Interest rate – 0.35, Current Interest Rate – Servicing Fee) * (Months between Last Principal & Interest paid to date and zero balance date) * 30/360/100. Please note that the following business rules are applied to this calculation: a. For all loans, 35 bps is used as a proxy for servicing fee, servicing fee will be used if higher than 35 bps. b. The Actual Loss Calculation will be set to zero for loans with a Defect Settlement Date value populated.	$ Amount	Numeric Literal Decimal

	c. The Actual Loss Calculation will be set to zero for loans with Net Sales Proceeds = 'U" (Net Sales Proceeds are missing), or expenses are not available. d. The Actual Loss Calculation will be set to missing for loans disposed within three months prior to the performance cutoff date. e. Modification Costs are currently not included in the calculation of the Actual Loss Calculation field.		
23	**MODIFICATION COST –** The cumulative modification cost amount calculated when Freddie Mac determines such mortgage loan has experienced a rate modification event or a UPB forbearance.. This amount will be calculated on a monthly basis beginning with the first reporting period a modification event is reported and disclosed in the last performance record. For example: calculate monthly modification cost as (min(Origination ANY, (Original Interest Rate - 0.35))/1200 * Current Actual UPB) – (min(Current ANY,(Current Interest Rate-0.35))/1200 * (Interest bearing upb)) ,and aggregate each month since modification through the Performance Cutoff Date into a cumulative amount. *For loans that go through a payment deferral program, cumulative modification cost includes interest foregone on the UPB deferred by the payment deferral until the last performance record.*	$ Amount	Numeric Literal Decimal
24	**STEP MODIFICATION FLAG** – A Y/N flag will be disclosed for every modified loan, to denote if the terms of modification agreement call for note rate to increase over time.	• Y = Current Period Step Mod • N = Current Period Non-Step Mod • Space (1) = Not a Current Period Mod	Alpha
25	**DEFERRED PAYMENT PLAN** – A flag will be disclosed to indicate a Deferred Payment Plan for the loan.	• Y = Current Period Deferred Plan • P = Prior Period Deferred Plan • Space (1) = No Deferred Plan	Alpha
26	**ESTIMATED LOAN TO VALUE (ELTV) –** A ratio indicating current LTV based on the estimated current value of the property obtained through Freddie Mac's Automated Valuation Model (AVM). Note: Only populated for April 2017 and following periods.	• 1 – 998 • 999 = Unknown • Blank = Data Not Available	Numeric Literal
27	**ZERO BALANCE REMOVAL UPB –** The amount of total UPB remaining on the loan immediately prior to the application of the Zero Balance Code	$ Amount	Numeric Literal Decimal
28	**DELINQUENT ACCRUED INTEREST** – The amount of delinquent interest owed by the borrower at the time of default. This field will only be populated for Zero Balance Codes 02, 03, 09, & 15. Note: the Delinquent Accrued Interest field will be set to zero for loans with a Defect Settlement Date value populated.	$ Amount	Numeric Literal Decima
29	**DELINQUENCY DUE TO DISASTER** – A flag indicating that the Servicer has reported disaster-related hardship as defined by the Freddie Mac Seller/Servicer Guide. Note: Only populated for January 2014 and following periods.	Y = Delinquency Due to Disaster	Alpha
30	**BORROWER ASSISTANCE STATUS CODE –** Regardless of delinquency status, the type of assistance plan that the borrower is enrolled in that provides temporary mortgage payment relief or an opportunity to cure a mortgage delinquency over a defined period. Note: Only populated for January 2014 and following periods.	F = Forbearance R = Repayment T = Trial Period	Alpha
31	**CURRENT MONTH MODIFICATION COST:** The current month modification cost amount calculated when Freddie Mac determines such mortgage loan has experienced a rate modification event or a UPB forbearance. This amount will be calculated on a monthly basis beginning with the first reporting	$ Amount	Numeric Literal Decimal

	period a modification event is reported and disclosed until the last performance record. For example: calculate monthly modification cost as (min(Origination ANY, (Original Interest Rate - 0.35))/1200 * Current Actual UPB) – (min(Current ANY,(Current Interest Rate-0.35))/1200 * (Interest bearing upb)) . For a loan that is acquired by Freddie Mac Homesteps and subsequently disposed as REO, Current Month Modification Cost represents the monthly modification cost aggregated during the period between REO acquisition and REO disposition. For loans that go through a payment deferral program, modification cost is calculated as interest foregone on the UPB deferred by the payment deferral.		
32	**INTEREST BEARING UPB :** The current interest bearing UPB of the modified mortgage.	$ Amount	Numeric Literal Decimal

6.2 EXPLORATORY ANALYSIS

Before any further, our first step is to exclude the variables that don´t give information to our model.

- **Origination Data file**: "MSA", "NUMBER OF UNITS", "CHANNEL", "PROPERTY STATE", "PROPERTY TYPE", "POSTAL CODE", "SELLER NAME", "SERVICER NAME", "PROGRAM INDICATOR", "RELIEF REFINANCE INDICATOR", "PROPERTY VALUATION METHOD"

- **Monthly Performance Data File**: "NET SALE PROCEEDS", "STEP MODIFICATION FLAG", "ESTIMATED LOAN TO VALUE", "DELINQUENT ACCRUED INTEREST", "DELINQUENCY DUE TO DISASTER", "CURRENT MONTH MODIFICATION COST", "INTEREST BEARING UPB", "BORROWER ASSISTANCE STATUS CODE"

Since this file contains monthly loan performance information, the Loan ID is not unique and may be repeated within a year. Thus, we need to make sure the Loan ID variable is identified as our next step. Having said that, as this would be the loan's ultimate result, we only chose the Loan ID with the greatest Loan age value. Three scenarios are feasible. There are three possible outcomes: the loan is in default, it is not in default, or it has been fully paid off.

Subsequently, we join the two sets of data, Origination and Performance ones by our Unique variable, the Loan ID.

6.2.1 MISSING VALUES

Predictive models may become biased and skew when variables with a significant percentage of missing data are included. This might result in forecasts that are unreliable and erroneous. Analysts can improve the accuracy and reliability of the model by ensuring that its predictions are grounded in complete and reliable information by eliminating such variables. Furthermore, plotting missing values allows analysts to assess the impact of missing data on model interpretability. Variables with high missingness may obscure the true relationships between predictors and outcomes, undermining the interpretability of the model.

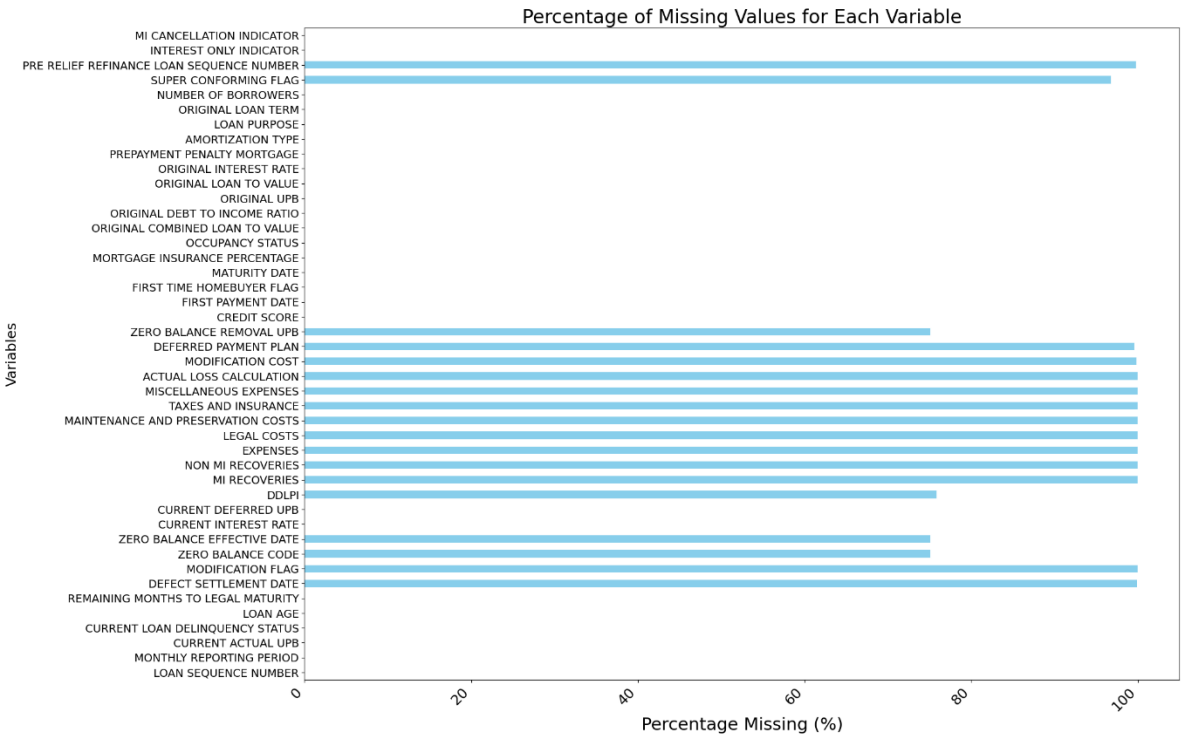


Figure 10 *Percentage of Missing Values for Feature*

Source: Authors Preparation

The variables are shown on the y-axis of Figure 10, while the volume of the missing elements is shown on the x-axis. Almost eighteen variables have more of 80% of missing values, so we will not include these variables in our research, regardless of whether they were absent at random because they cannot be imputed.

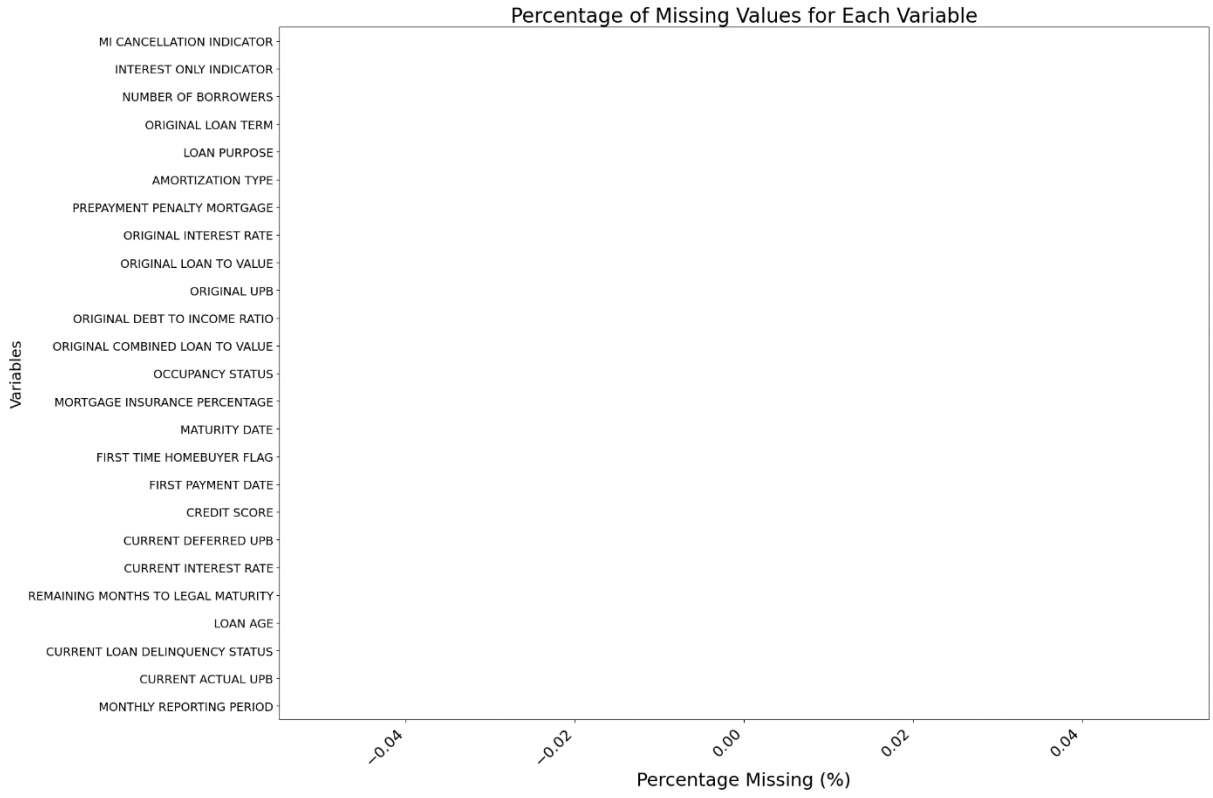


Figure 11 *Percentage of Missing values after removing the features*

Source: Authors Preparation

6.2.2 CREATION OF DEFAULT (TARGET) VARIABLE

Using the data's current loan delinquent state, we computed default. Delinquency status is a Freddie Mac-specified metric that counts the days that a borrower has fallen behind on their monthly commitments. Status 0 in the data denotes current or less than 30 days, Status 1 indicates more than 30 days but less than 60 days, Status 2 indicates more than 60 days but less than 90 days, and Status 3 denotes delinquent for days between 90 and 119. For every loan, we determined the highest delinquent state, and we then specified the output variable as follows: if the maximum delinquency status is greater than 3 (i.e. loan is of delinquency status 3 and above) or zero balance code is greater than 9 (zero balance code is used to indicate why the loan balance was reduced to zero, 09 indicates deed in lieu loans) we classify such loan as default.

After computing and counting the new variable, we have 247238 with value = 0 (non-default loan) and 2743 with value = 1 (default loan) which tends to be common for data related to mortgages, as normally the last credit to be given up by the client is the house credit.

6.2.3 DATA TYPE

Many machine learning algorithms require numeric input. By converting all variables to numeric or float format, we ensure that our data is compatible with a wide range of modeling techniques. Algorithms such as linear regression and decision trees operate on numeric data, and failing to convert variables appropriately may lead to errors or inaccurate results.

```
print(merged_table.dtypes)

MONTHLY REPORTING PERIOD              float64
CURRENT ACTUAL UPB                    float64
CURRENT LOAN DELINQUENCY STATUS       float64
LOAN AGE                              float64
REMAINING MONTHS TO LEGAL MATURITY    float64
CURRENT INTEREST RATE                 float64
CURRENT DEFERRED UPB                  float64
CREDIT SCORE                            int64
FIRST PAYMENT DATE                      int64
FIRST TIME HOMEBUYER FLAG              object
MATURITY DATE                           int64
MORTGAGE INSURANCE PERCENTAGE           int64
OCCUPANCY STATUS                       object
ORIGINAL COMBINED LOAN TO VALUE         int64
ORIGINAL DEBT TO INCOME RATIO           int64
ORIGINAL UPB                            int64
ORIGINAL LOAN TO VALUE                  int64
ORIGINAL INTEREST RATE                float64
PREPAYMENT PENALTY MORTGAGE            object
AMORTIZATION TYPE                      object
LOAN PURPOSE                           object
ORIGINAL LOAN TERM                      int64
NUMBER OF BORROWERS                     int64
INTEREST ONLY INDICATOR                object
MI CANCELLATION INDICATOR              object
STATUS                                  int64
```

Figure 12 *Data Type of Each Feature*

Source: Authors Preparation

As we can see in Figure 12 we have seven variables non numeric: “FIRST TIME HOMEBUYER FLAG”, “OCCUPANCY STATUS”, “PREPAYMENT PENALTY MORTAGE”, “AMORTIZATION TYPE”, “LOAN PURPOSE”, “INTEREST ONLY INDICATOR”, “MI CANCELLATION INDICATOR”. It´s easy to transform this variables into integers, as most of them only take two values, so it is easy to match them with numbers such as “1” and “2” or “0” if it`s non value.

6.3 TRAIN DATA AND TEST DATA

It is vital to make sure that an appropriate assessment design is employed in addition to choosing suitable performance measures to assess trained models. The hold-out method was selected as the assessment design for this thesis. The hold-out method for model evaluation represents the mechanism of splitting the dataset into training and test datasets. The model is trained on the training set and then tested on the testing set to get the most optimal model.

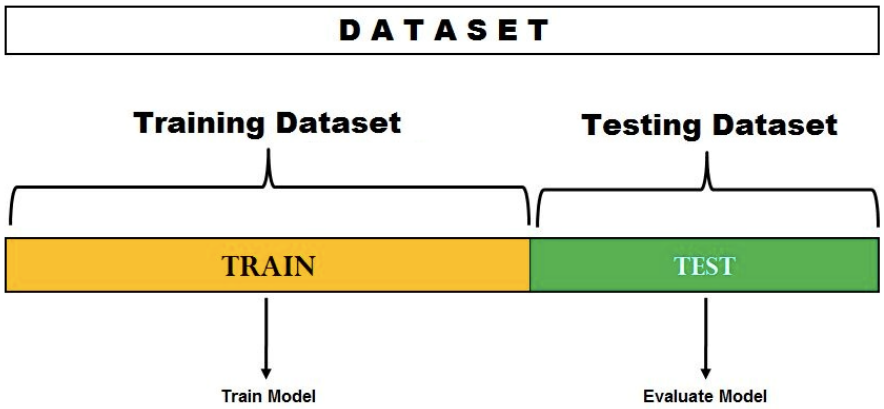


Figure 13 *Hold-out method for model evaluation*

Source: Analytics Yogi

The data was split in two datasets:

- Training dataset, with 80% of the data;
- Test dataset, with 20% of the data.

A machine learning model's training dataset is an essential part of the process. Usually comprising 80% of the original data, as in this study, it comprises a substantial amount of the data. The training dataset serves the main function of giving the model a large collection of instances to work with and learn from. In order to reduce prediction errors, the model modifies its internal parameters throughout the training phase by analyzing the input variables and the related target variables. Finding patterns, connections, and trends in the data is part of this learning process that helps the model produce precise predictions. The

model can perform better on unknown data and achieve better generalization by utilizing a significant chunk of the input.

The remaining 20% of the original data, known as the test dataset, has an equally significant but distinct function in the modeling process. Its main purpose is to assess the trained model's performance and generalizability. The model is evaluated on this different dataset, which it did not come into contact with during training, after it has completed its training. Predictions are made based on the test data and compared to the actual target values as part of this evaluation. The results provide an unbiased estimate of how well the model is likely to perform on new, unseen data. By using a dedicated test dataset, we can assess the model's accuracy, detect any overfitting, and ensure that it generalizes well to real-world scenarios. This step is critical for validating the model's reliability and effectiveness before deploying it for practical use.

6.4 DATA PROCESSING

6.4.1 OUTLIERS

The practice of identifying data points whose behavior deviates significantly from expectations is known as outlier detection. A single data point that deviates significantly from the other data points in the set is called an outlier. Although there are several reasons why this dataset anomaly could arise, outliers frequently happen as a result of human mistake in data collection, manipulation, or processing. Finding outliers is a crucial part of data preparation since they have the potential to distort general data trends.

The outliers were detected and removed through the IQR method. For this matter the q3 (75th percentile) and q1 (25th percentile) were computed in order to obtain the inter quartile range (q3- q1), then it was possible to calculate the lower and upper bound:

Lower Bound = q1–1.5*IQR

Upper Bound = q3+1.5*IQR

Any value below the lower bound and above the upper bound are considered to be outliers.

6.4.2 DROPPING HIGH CORRELATED FEATURES

To evaluate the monotonic relationship between characteristics and the target variable, as well as between features and each other, Spearman correlation analysis was used. We are not limited by assumptions about the distributions of the variables when assessing the direction and strength of the link thanks to this non-parametric correlation measure. It is also especially useful for investigating correlations between continuous and ordinal data.

Figure 14 *Correlation Matrix*

Source: Authors Preparation

Based on our predefined criteria (values >= 0.7 or <= -0.7), we found the following highly correlated variables and constant variables (variables with no value in Figure 14). They were removed from our train and test data: "ORIGINAL LOAN TERM","CURRENT DEFERRED UPB","ORIGINAL LOAN TO VALUE","CURRENT LOAN DELINQUENCY STATUS","MATURITY DATE", "ORIGINAL INTEREST RATE", "PREPAYMENT PENALTY MORTGAGE", "AMORTIZATION TYPE", "INTEREST ONLY INDICATOR".

These correlations indicate consistent associations between the variables, suggesting that changes in one variable are intricately linked to corresponding changes in the other variable.

6.4.3 Features Importance

To further select valuable variables, we employed Random Forests, which naturally ranks features based on importance due to their tree-like structure. Random Forest Classifier was utilized with default parameters, allowing the trees to expand fully. We then fitted it to the training data and plotted the feature importance's calculated by the model implicitly.

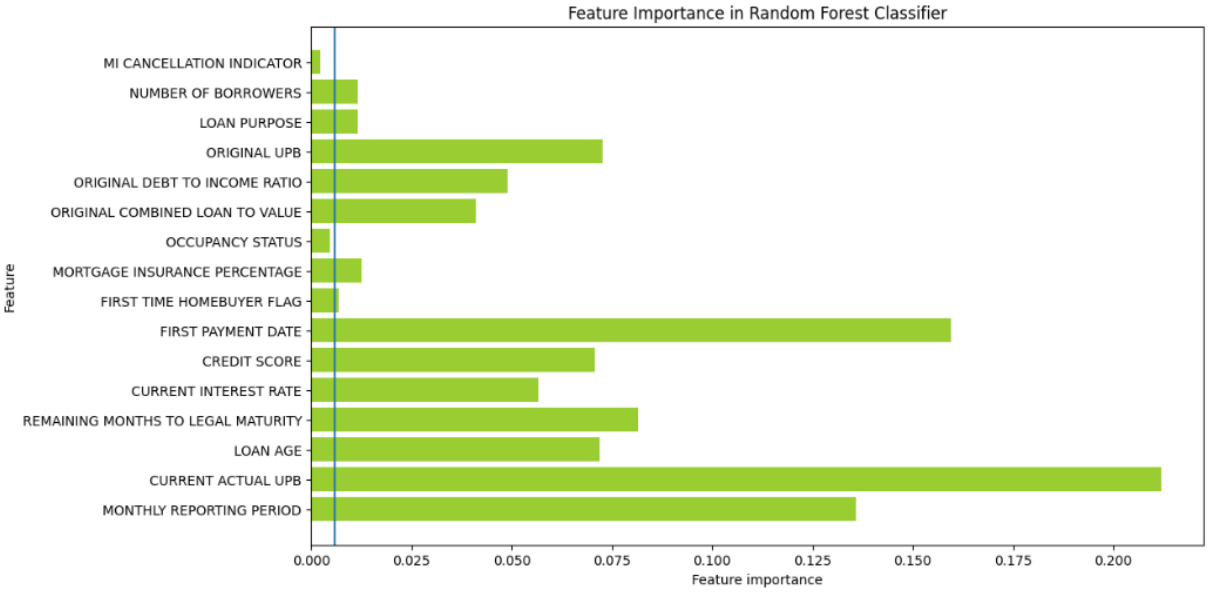


Figure 15 *Feature Importance*

Source: Authors Preparation

By looking at the bar chart above, the features that were situated on the left side of the line (<0.006) were removed and the features with high importance, but with little meaning for the business model were removed too ("MI CANCELLATION INDICATOR","FIRST PAYMENT DATE","MONTHLY REPORTING PERIOD","OCCUPANCY STATUS","FIRST TIME HOMEBUYER FLAG") .

6.5 OVERSAMPLING

One issue with the Freddie Mac loan level dataset is highly unbalanced distribution of the two classes default and non-default. 99% of the observations were defined as non-default while just 1% of the data is assigned to class 1 (defaulted). In this case the classifiers won't be able to recognize minor classes and are influenced by major classes.

Data that is not balanced is problematic. This is because classifiers tend to do badly on the minority class because of their bias towards the majority class as a result of the class difference. Therefore, it is required to balance the data before fitting the model over the training dataset and forecasting classes over the testing dataset. To try to solve this issue, a number of methods have been put out up to this point. A few of them try to resample the data by adding records from the minority class or undersampling, which eliminates records from the majority class. Ensuring that each target class is represented in an equal proportion is the primary objective of the suggested solutions.

In this thesis will be used SMOTE (Hu & Li, 2013) (Synthetic Minority Oversampling Technique). This algorithm is capable of overcoming class inbalancement in a more robust way than just by simply performing oversampling or undersampling. Because in this case, it is not necessary

to remove observations from the majority class (undersampling) which would imply losing information. Thus, is also not necessary to create duplicate observations of the minority class (oversampling) which could later translate into a problem with overfitting. This algorithm is able to generate artificial samples based on the linear interpolation of both classes.

After using SMOTE, the number of Target values with non-default value (0) is equal to default value (1).

6.5.1 DROPPING HIGH CORRELATED FEATURES AFTER OVERSAMPLING

Oversampling is a useful strategy for correcting class imbalance, however after using these methods, it's critical to keep an eye on the relationships between the features because as we had few default cases, we had to add a lot of cases and it may have messed up the data. For the purpose of creating reliable and broadly applicable models, it is imperative to take measures to reduce the impact of increased correlations on model performance and interpretability.

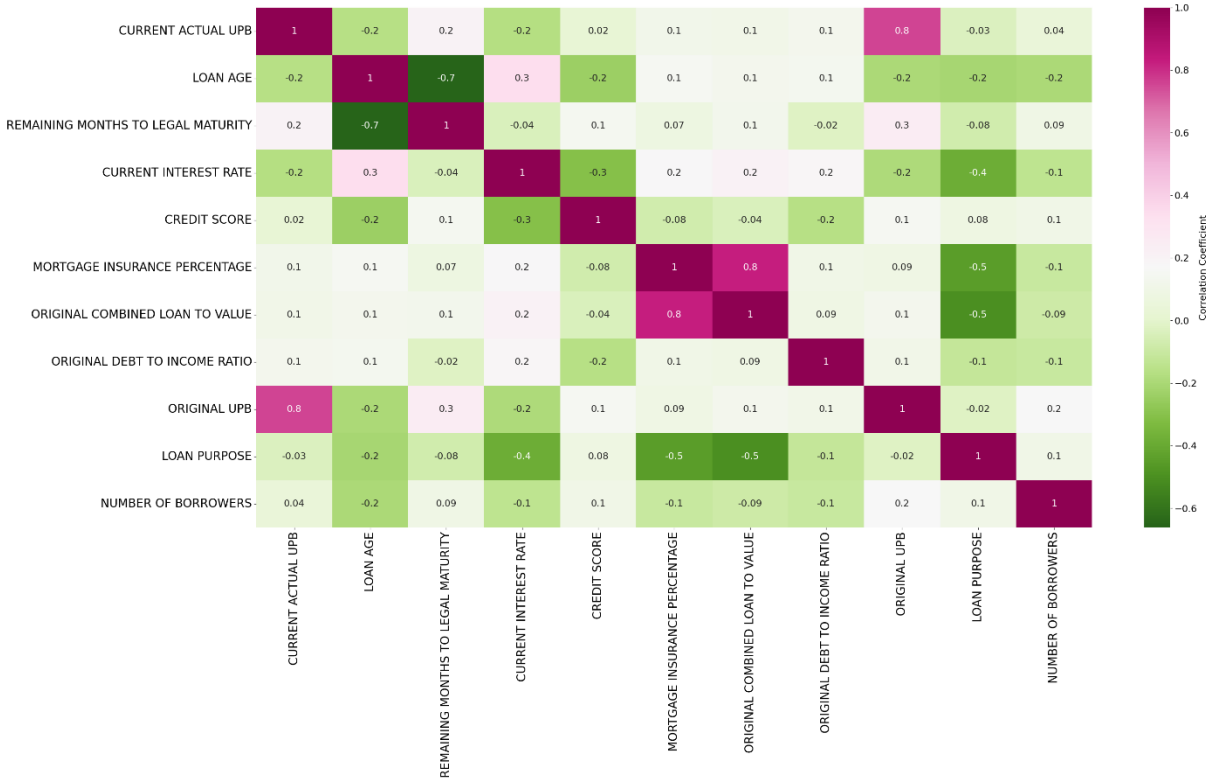


Figure 16 *Correlatiom Matrix after Oversampling*

Source: Authors Preparation

As we have already carried out a correlation study, in this one we have increased the criterion to >=0.8 and <=-0.8 to ensure that we only eliminate very highly correlated variables to minimize the loss of information from our data. The variables were: "ORIGINAL COMBINED LOAN TO VALUE","ORIGINAL UPB".

7. RESULTS AND DISCUSSION

In this section, we will be presenting the results of our classification models.

7.1 TYPE I ERROR

With a Type I error of 0.0068, XGBoost achieved the best performance in terms of good clients being mistakenly labeled as bad. Random Forest trailed closely after, with a 0.0105 inaccuracy. Logistic Regression had the highest Type I error at 0.1972, while Gradient Boosting showed the highest Type I error at 0.0568.

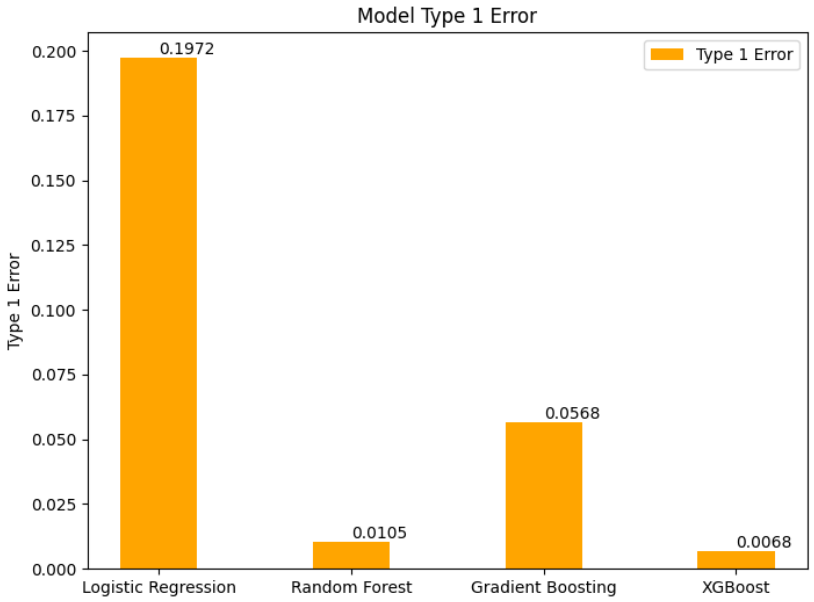


Figure 17 *Models Type I Error*

Source: Authors Preparation

7.2 TYPE II ERROR

With 0.0057, Random Forest had the lowest error for the metric Type II error (bad customers misclassified as good), followed by 0.0104 for XGBoost. The largest Type II error was 0.1892 for Logistic Regression while the highest error for Gradient Boosting was 0.0484.

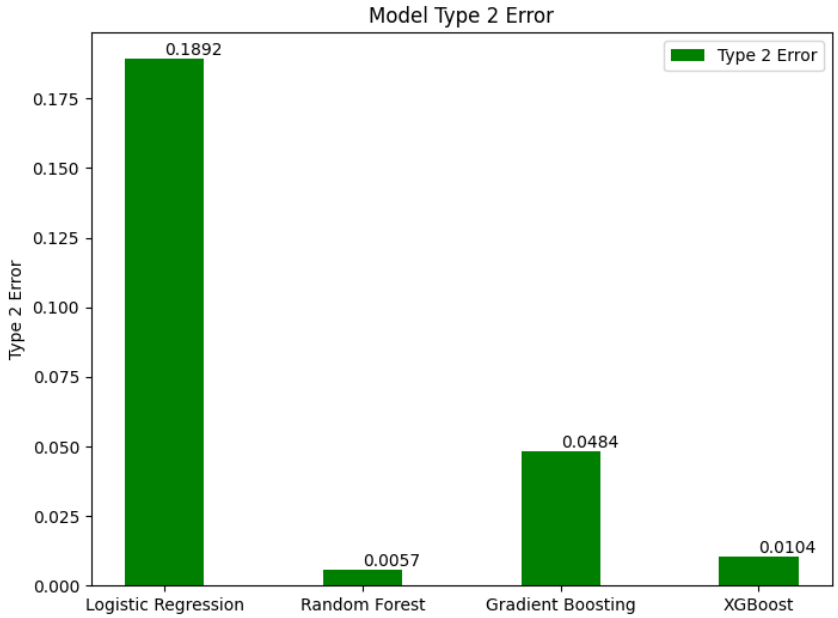


Figure 18 *Models Type II Error*

Source: Authors Preparation

7.3 Accuracy

When evaluating the predictive capacity of the model, the accuracy of the fraction of correctly classified loans (good and bad), the best model is XGBoost with 0.9532, in the second place with 0.9474 is the Gradient Boosting, Random Forest and Logistic Regression with 0.9120 and 0.8068 respectively.

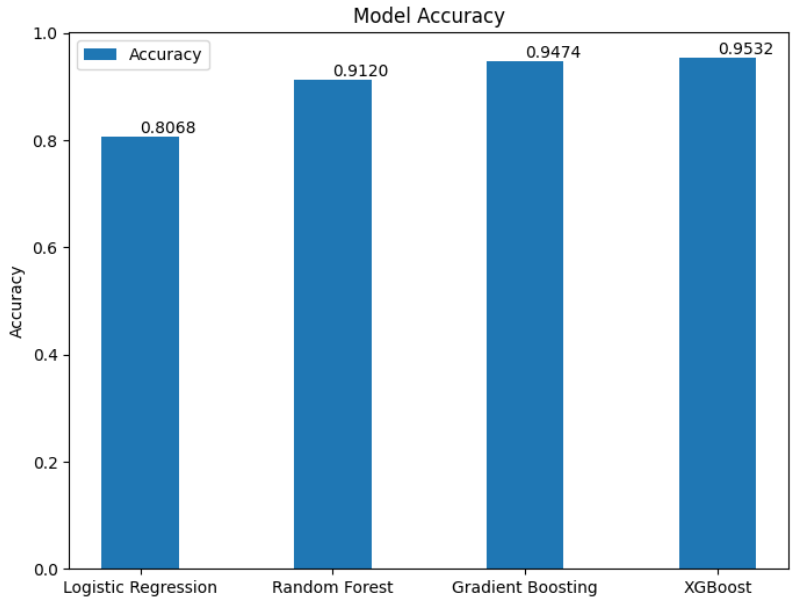


Figure 19 *Models Accuracy*

Source: Authors Preparation

7.4 AUC Score and ROC Curve

With an AUC value of 0.8837, Logistic Regression demonstrated a favorable outcome in the metric. Random Forest came in third with a 0.9120 value. Gradient Boosting received an even higher score of 0.9474, while XGBoost demonstrated the best performance with an AUC of 0.9532.

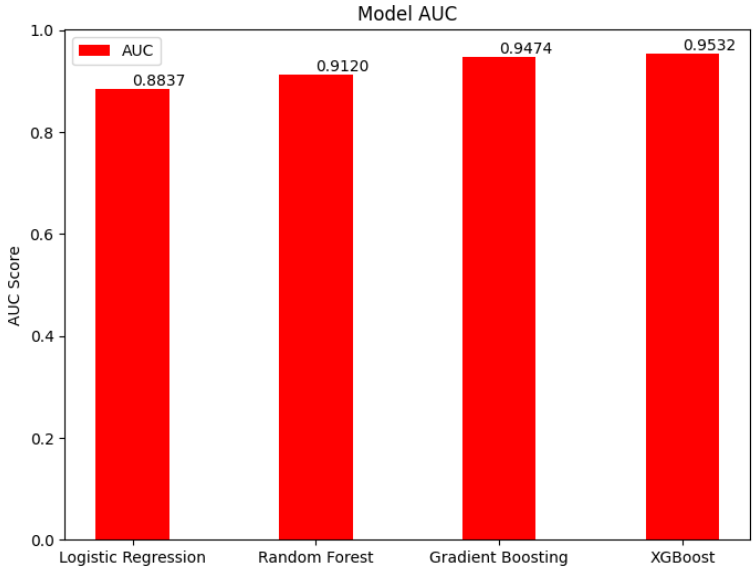


Figure 20 *Models AUC Scorre*

Source: Authors Preparation

The ROC Curves for each model are shown by Figure 20. For various parameter threshold points, the true positive rate is represented by the horizontal axes of a ROC curve, and the false positive rate is represented by the vertical axes. The forecast is therefore more accurate if the curve is closer to the upper left.

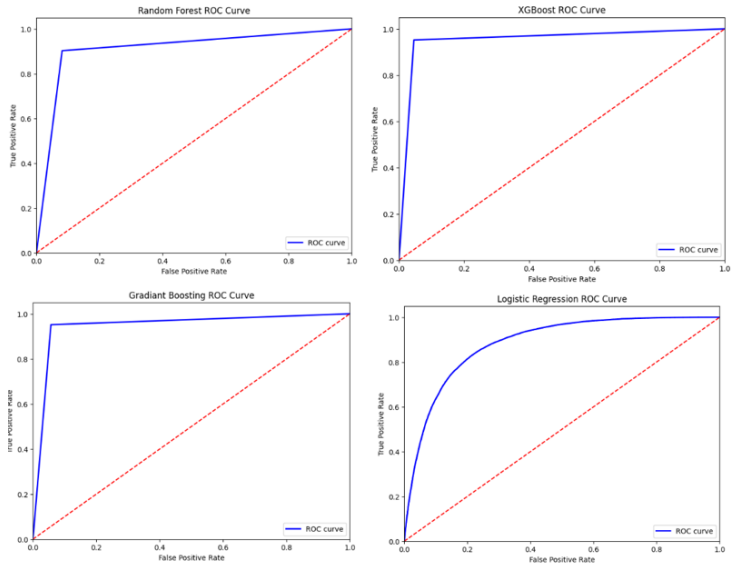


Figure 21 *Models ROC curves*

Source: Authors Preparation

7.5 SUMMARY RESULTS AND DISCUSSION

In this thesis, we aim to compare the effectiveness of many ensemble learning techniques, including Gradient Boosting, Random Forest, XGBoost, and the industry standard, Logistic Regression. The criteria selected for model evaluation are crucial because they affect the comparability and measurement of machine learning algorithm performance. Accuracy, AUC, type I and type II errors were the four performance indicator measures selected in order to validate our models and arrive at a dependable and solid conclusion.

Firstly, in terms of accuracy, the models generally demonstrate high performance. XGBoost leads the pack with an accuracy of 0.9532, closely followed by Gradient Boosting at 0.9474. Random Forest also performs well with an accuracy of 0.9120, while Logistic Regression trails slightly behind with 0.8068.

Random Forest and XGBoost achieve the lowest errors when Type I error (misclassifying good customers as bad) is taken into account, demonstrating their efficacy in detecting positive cases. The next model with a somewhat low error is gradient boosting, and the model with the largest Type I error is logistic regression.

Nonetheless, Random Forest and XGBoost fare remarkably well in Type II mistake (misclassifying bad clients as good), demonstrating their capacity to recognize negative situations with accuracy. Here, their error count is the lowest. While Gradient Boosting performs exceptionally well, Logistic Regression exhibits the largest Type II error, suggesting that it struggles to accurately detect negative events.

As the last metric to justify the model's performance in terms of separation between positive and negative cases, the AUC score shows that the XGBoost outperforms the other models. It has an AUC score of 0.9532. With an excellent discriminating power, 0.9474, the Gradient Boosting ranks second. The Random Forest has a fairly good AUC value of 0.9120; whereas, the Logistic Regression has a slightly inferior discriminating power, 0.8837.

Logistic Regression is pretty accurate overall, but it tends to misclassify many good customers as bad ones (high Type I errors). On the other hand, Random Forest is good at correctly identifying good customers, but its overall accuracy isn't the highest. This shows why it's important to use a variety of metrics when evaluating credit scoring models, so we get a complete picture of their performance.

The overall performance of the XGBoost is the best. It has good discriminating power, low error rates and high accuracy with respect to majority of the criteria. Logistic Regression, though competitive, has certain limitations especially in terms of error rates and discriminating power; on the other hand, the Random Forest and Gradient Boosting perform exceptionally well overall.

8. CONCLUSION

This study started off by stating that credit scoring is one of the most important problems in financial industry especially for banks in granting mortgage applications. For many years logistic regression has served as a gold standard due to its simplicity and easy interpretation by regulators and credit risk managers alike. But with the increasing use of big data, new advanced models have been devised and are outperforming logistic regression.

Our study compared logistic regression with some ensemble methods such as Gradient Boosting, Random Forest and XGBoost. It was evident from the results that ensemble methods outperform logistic regression in many respects. The advanced models provided higher accuracy and higher AUC scores (especially for gradient boosting) in predicting the creditworthiness of applicants thereby providing much better discrimination between good and bad applicants.

The results show that sophisticated ensemble methods used for the credit risk prediction outperform the traditional logistic regression approach. Therefore, credit risk institutions can improve their credit scoring practices by adopting more sophisticated ensemble methods. On the other hand, this might not be acceptable in some cases from a regulatory point of view. Credit risk institutions need to find the balance between the regulatory constraints and the increase in the predictive accuracy. This study shows that credit scoring should continually evolve in order to handle the financial credit risk and to make accurate lending decisions.

Regarding the research questions: “Will machine learning models be better than Logistic Regression? “Is the model that accepts more credit the safest for the bank?”

For the first question, machine learning models such as Random Forest, Gradient Boosting, and XGBoost not only provide higher accuracy and AUC scores but also significantly reduce the rates of Type I and Type II errors compared to logistic regression. Therefore, machine learning models can be considered better than logistic regression for credit scoring tasks based on these metrics

For the second one, we must compare the Type I error rates (i.e., false positives) of the two models. The Type I error rate is the proportion (or percentage) of good customers who are mistakenly classified as bad customers. The smaller the Type I error rate, the fewer good customers are mistakenly rejected, and the more credit the bank will extend.

XGBoost exhibits the highest accuracy and AUC score, so it is the overall best in separating good from bad customers. Also, XGBoost exhibits the lowest Type I error, and will reject fewer good customers, which is good for approving more credit. XGBoost does exhibit a slightly higher Type II error (0.0104) than Random Forest (0.0057), which means it will misclassify slightly more bad customers as good than Random Forest.

XGBoost will likely approve more credit and generally performs the best on most metrics, but Random Forest has a lower Type II error, which means it is slightly better at not misclassifying bad customers as good, so if the main concern is the risk of default (Type II errors), Random Forest is probably safer. But overall, XGBoost performs the best and is probably safe too, especially with its high accuracy and low Type I error, so the ans.

9. LIMITATIONS AND RECOMMENDATIONS

One of the main limitations found for this thesis was the long time needed to run some models, given the high volume of data. This constraint influenced the overall efficiency and feasibility of the modeling process.

A major suggestion to better the data-splitting method is that it could partition the dataset into three subsets: training, validation, and test sets, instead of the normal practice of splitting the dataset into two subsets of training and test sets. This approach has several important benefits: overfitting can be prevented because the test sets are taken to be independent. The process of iterative training-validation-testing makes way for an ever-improving model. The performance obtained through the validation and the changes that need to be incorporated can be modified and improved in the model, and this can be tested back on the validation set, time and again, until an improved model is obtained.

The second recommendation is to add hyperparameter tuning in future research. Through hyperparameter optimization, we will ensure that the machine learning models are fine-tuned to attain maximum performance that delivers forecasts with more accuracy and reliability. Eventually, this strategy will enhance the overall effectiveness of the credit scoring models by helping to develop robust models that generalize well to new, unseen data.

Finally, the addition of features with full macroeconomic data into the process of building the credit scoring model, such as home prices or the rate of unemployment, will add strength to the model's interpretation and accuracy.

10. BIBLIOGRAPHY

Addo, P. M., Guegan, D., & Hassani, B. (2018). *Credit risk analysis using machine and deep learning models. Risks, 6 (2), 1–20*.

Ala'raj, M., & Abbod, M. F. (2016). Classifiers consensus system approach for credit scoring. *Knowledge-Based Systems*, *104*, 89–105.

Ashofteh, A., & Bravo, J. M. (2021). A conservative approach for online credit scoring. *Expert Systems with Applications*, *176*, 114835.

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADS-DM*.

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*, 1937–1967.

Bravo, J. M. (2020). *Longevity-linked life annuities: A Bayesian model ensemble pricing approach*.

Bravo, J. M., & Ashofteh, A. (2023). *Ensemble Methods for Consumer Price Inflation Forecasting*.

Bravo, J. M., & Ayuso, M. (2021). *Forecasting the retirement age: A Bayesian model ensemble approach*. 123–135.

Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W., & Siddique, A. (2016). Risk and risk management in the credit card industry. *Journal of Banking & Finance*, *72*, 218–239.

Chamboko, R., & Bravo, J. M. (2016). On the modelling of prognosis from delinquency to normal performance on retail consumer loans. *Risk Management*, *18*, 264–287.

Chamboko, R., & Bravo, J. M. (2019). Modelling and forecasting recurrent recovery events on consumer loans. *International Journal of Applied Decision Sciences*, *12*(3), 271–287.

Chamboko, R., & Bravo, J. M. (2020). A multi-state approach to modelling intermediate events and multiple mortgage loan outcomes. *Risks*, *8*(2), 64.

Chen, T., & Guestrin, C. (2016). *Xgboost: A scalable tree boosting system*. 785–794.

Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, *183*(3), 1447–1465.

Dornigg, T. (2022). *Credit risk modeling-predicting customer loan defaults with machine learning models*.

Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, *297*(3), 1178–1192.

Durand, D. (1941). *Risk elements in consumer instalment financing* (Números dura41-1). National bureau of economic research.

Engelmann, B., & Rauhmeier, R. (2006). *The Basel II risk parameters: Estimation, validation, and stress testing*. Springer Science & Business Media.

Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. *Machine learning*, *31*(1), 1–38.

Fensterstock, A. (2005). Credit scoring and the next step. *Business credit*, *107*(3), 46–49.

Fitzpatrick, T., & Mues, C. (2016). An empirical comparison of classification algorithms for mortgage default prediction: Evidence from a distressed mortgage market. *European Journal of Operational Research*, *249*(2), 427–439.

Hu, F., & Li, H. (2013). A novel boundary oversampling algorithm based on neighborhood rough set model: NRSBoundary-SMOTE. *Mathematical Problems in Engineering*, *2013*, 1–10.

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, *34*(11), 2767–2787.

Koh, H. C., Tan, W. C., & Goh, C. P. (2006). A two-step method to construct credit scoring models with data mining techniques. *International Journal of Business and Information*, *1*(1), 96–118.

Li, X.-L., & Zhong, Y. (2012). *An overview of personal credit scoring: Techniques and future work*.

Lin, F. Y., & McClean, S. (2001). A data mining approach to the prediction of corporate failure. *Knowledge-based systems*, *14*(3–4), 189–195.

Liu, J., Zhang, S., & Fan, H. (2022). A two-stage hybrid credit risk prediction model based on XGBoost and graph-based deep neural network. *Expert Systems with Applications*, *195*, 116624.

Mac, F. (2019). Single family loan-level dataset general user guide. *Freddie Mac. Available online: http://www. freddiemac. com/fmac-resources/research/pdf/user_guide. pdf (accessed on 3 February 2019)*.

Maheswari, P., & Narayana, C. V. (2020). *Predictions of loan defaulter-A data science perspective*. 1–4.

Mamonov, S., & Benbunan-Fich, R. (2017). What can we learn from past mistakes? Lessons from data mining the Fannie Mae mortgage portfolio. *Journal of real estate research*, *39*(2), 235–262.

Mester, L. J. (1997). What's the point of credit scoring. *Business review*, *3*(Sep/Oct), 3–16.

Mhammedi, S., El Massari, H., Gherabi, N., & Amnai, M. (2023). *Enhancing Book Recommendations on GoodReads: A Data Mining Approach Based Random Forest Classification*. 395–409.

Mohammed, A., & Kora, R. (2023). A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University-Computer and Information Sciences*, *35*(2), 757–774.

Nalluri, M., Pentela, M., & Eluri, N. R. (2020). A Scalable Tree Boosting System: XG Boost. *Int. J. Res. Stud. Sci. Eng. Technol*, *7*, 36–51.

Nariya, M. K., Mills, C. E., Sorger, P. K., & Sokolov, A. (2023). Paired evaluation of machine-learning models characterizes effects of confounders and outliers. *Patterns*, *4*(8).

Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision support systems*, *50*(3), 559–569.

Orgler, Y. E. (1970). A credit scoring model for commercial loans. *Journal of money, Credit and Banking*, *2*(4), 435–445.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and systems magazine*, *6*(3), 21–45.

Qiu, W. (2019). *Credit risk prediction in an imbalanced social lending environment based on XGBoost*. 150–156.

Raimundo, B., & Bravo, J. M. (2023). *Credit Risk Scoring: A Stacking Generalization Approach*. 382–396.

Sealand, J. C. (2018). Short-term prediction of mortgage default using ensembled machine learning models. *Unpublished doctoral dissertation]. Slippery Rock University*.

Sheikh, M. A., Goel, A. K., & Kumar, T. (2020). *An approach for prediction of loan approval using machine learning algorithm*. 490–494.

Spackman, K. A. (1989). *Signal detection theory: Valuable tools for evaluating inductive learning*. 160–163.

Trivedi, S. K. (2020). A study on credit scoring modeling with different feature selection and machine learning approaches. *Technology in Society*, *63*, 101413.

Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert systems with applications*, *38*(1), 223–230.

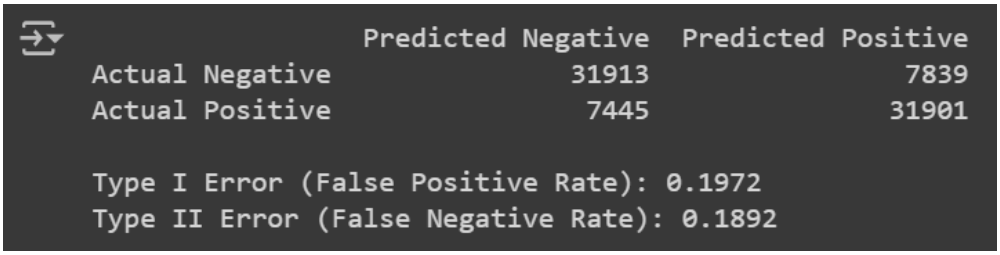
Wang, K., Li, M., Cheng, J., Zhou, X., & Li, G. (2022). Research on personal credit risk evaluation based on XGBoost. *Procedia computer science*, *199*, 1128–1135.

Xia, Y., Zhao, J., He, L., Li, Y., & Niu, M. (2020). A novel tree-based dynamic heterogeneous ensemble method for credit scoring. *Expert Systems with Applications*, *159*, 113615.

Zhou, Z.-H. (2012). *Ensemble methods: Foundations and algorithms*. CRC press.

APPENDIX

Figures A

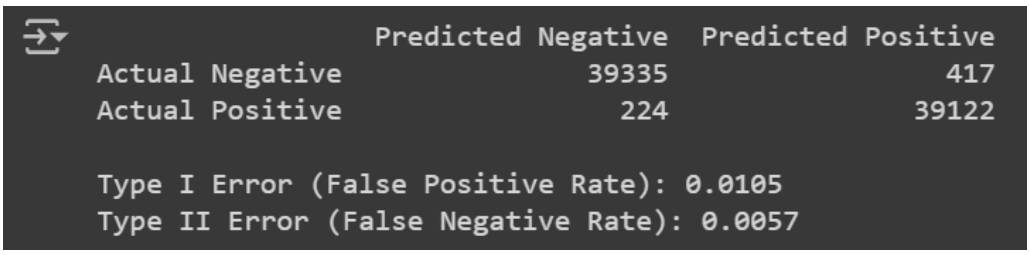


	Predicted Negative	Predicted Positive
Actual Negative	31913	7839
Actual Positive	7445	31901

Type I Error (False Positive Rate): 0.1972
Type II Error (False Negative Rate): 0.1892

A1 *Logistic Regression Resume*

Source: Authors Preparation

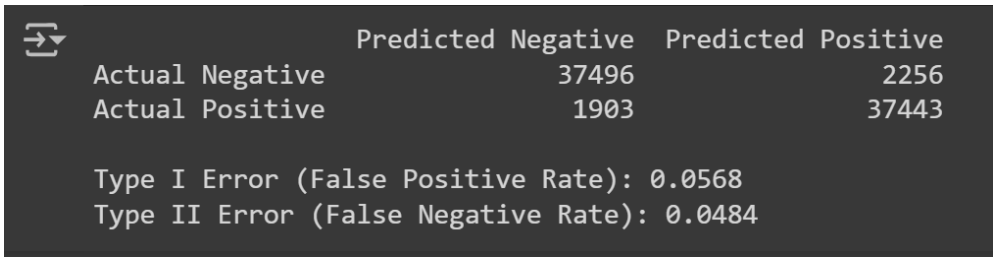


	Predicted Negative	Predicted Positive
Actual Negative	39335	417
Actual Positive	224	39122

Type I Error (False Positive Rate): 0.0105
Type II Error (False Negative Rate): 0.0057

Figure A2 *Random Forest Resume*

Source: Authors Preparation



	Predicted Negative	Predicted Positive
Actual Negative	37496	2256
Actual Positive	1903	37443

Type I Error (False Positive Rate): 0.0568
Type II Error (False Negative Rate): 0.0484

Figure A3 *Gradient Boosting Resume*

Source: Authors Preparation

	Predicted Negative	Predicted Positive
Actual Negative	39480	272
Actual Positive	408	38938

Type I Error (False Positive Rate): 0.0068
Type II Error (False Negative Rate): 0.0104

Figure A4 *XGBoost Resume*

Source: Authors Preparation

	CURRENT ACTUAL UPB	LOAN AGE	REMAINING MONTHS TO LEGAL MATURITY
count	316388.000000	316388.000000	316388.000000
mean	225306.830684	20.334719	328.924243
std	157246.933370	12.844115	26.491114
min	0.000000	0.000000	271.500000
25%	115042.407603	9.000000	320.000000
50%	205658.894681	19.000000	335.000000
75%	322020.970525	31.266769	349.000000
max	757500.000000	42.500000	403.500000

	CURRENT INTEREST RATE	CREDIT SCORE	MORTGAGE INSURANCE PERCENTAGE
count	316388.000000	316388.000000	316388.000000
mean	4.321695	733.214480	8.913865
std	0.928766	45.003839	11.382521
min	1.500000	617.000000	0.000000
25%	3.625000	700.000000	0.000000
50%	4.486835	734.000000	0.000000
75%	4.929334	770.000000	20.000000
max	7.187500	889.000000	30.000000

	ORIGINAL DEBT TO INCOME RATIO	LOAN PURPOSE	NUMBER OF BORROWERS
count	316388.000000	316388.000000	316388.000000
mean	37.533105	1.535972	1.306447
std	8.648087	0.742448	0.474452
min	5.500000	1.000000	1.000000
25%	32.035007	1.000000	1.000000
50%	39.118796	1.000000	1.000000
75%	44.000000	2.000000	2.000000
max	65.500000	3.000000	5.000000

Figure A5 *Descriptive Statistics before Modelling*

Source: Authors Preparation

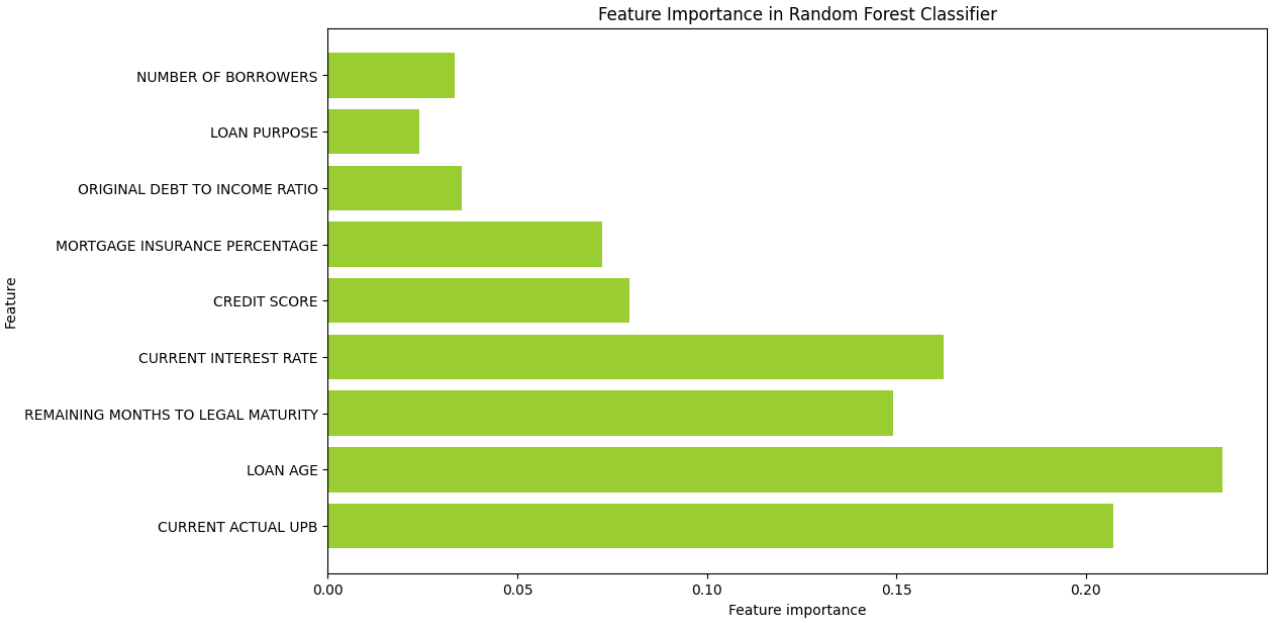


Figure A6 *Features Importance Before Modelling*

Source: Authors Preparation

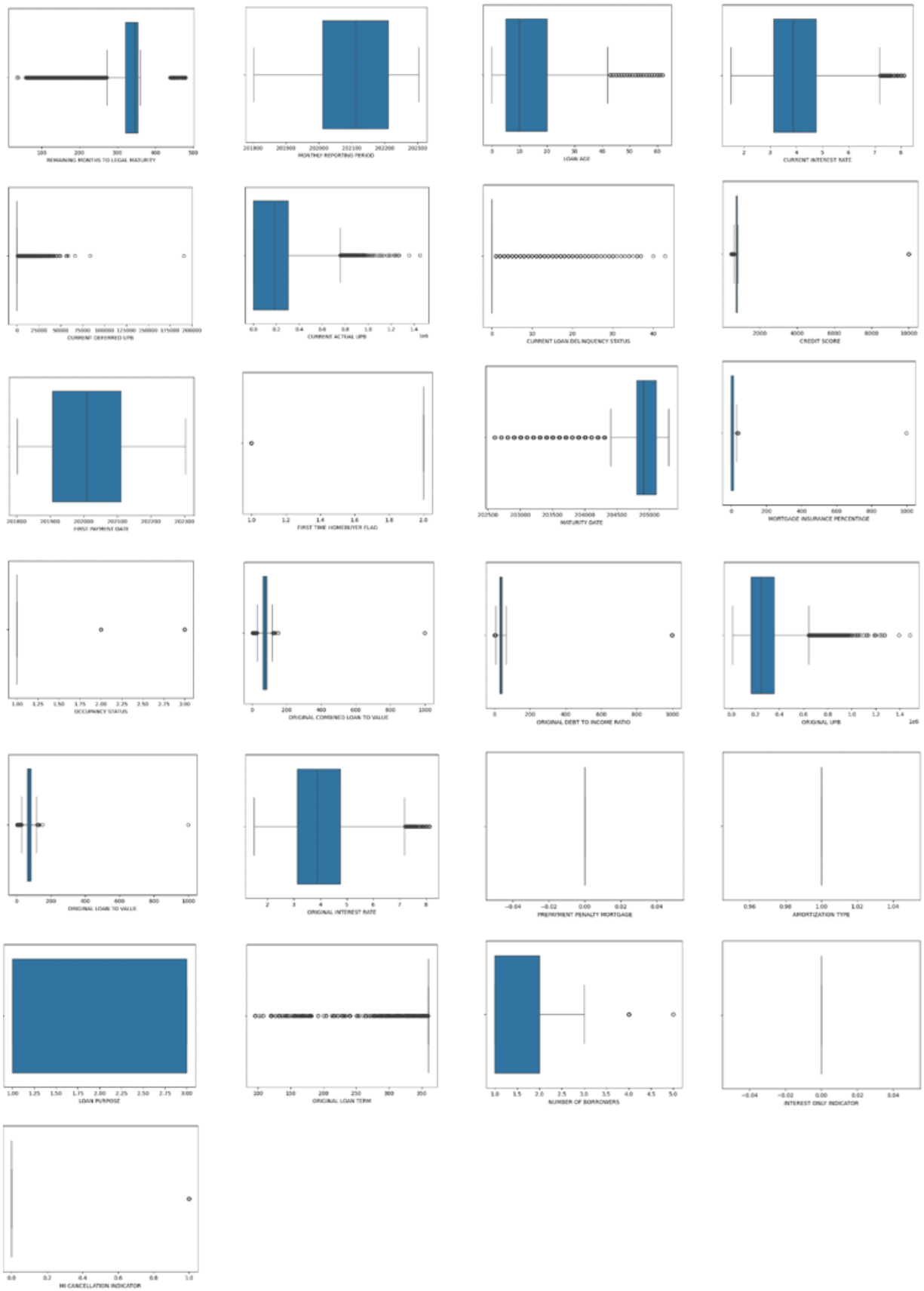


Figure A7 *Variables Box plots*

Source: Authors Preparation

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa