OPTIMIZING CREDIT RISK ASSESSMENT THROUGH COST-SENSITIVE LEARNING

BALANCING FALSE NEGATIVES, PROFIT MAXIMIZATION, AND REAL-WORLD SCENARIO WITH CLASS WEIGHTS IN PEER-TO-PEER LENDING

STAVROS CHATZIPAVLIDIS

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

u774845

COMMITTEE

dr. Grzegorz Chrupała
dr. Gonzalo Nápoles

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

January 22nd, 2024

WORD COUNT

8162

ACKNOWLEDGMENTS

I want to express my deepest gratitude to my mother, whose unwavering belief in me has been a constant source of motivation and strength throughout my academic journey. Her support and encouragement not only enabled me to pursue my dreams but also empowered me to navigate and overcome any challenges that came my way. I am also immensely thankful to my partner for her persistent support and for embarking on this adventure with me all the way to the Netherlands. Her presence and encouragement have been invaluable, making every step of this academic endeavor more meaningful. I extend my appreciation to my supervisor, Dr. Grzegorz Chrupała, for his exceptional guidance and insightful feedback throughout the process of writing this thesis. Additionally, I would like to thank Dr. Gonzalo Nápoles for graciously serving as the second reader for this thesis.

OPTIMIZING CREDIT RISK ASSESSMENT THROUGH COST-SENSITIVE LEARNING

BALANCING FALSE NEGATIVES, PROFIT MAXIMIZATION, AND REAL-WORLD SCENARIO WITH CLASS WEIGHTS IN PEER-TO-PEER LENDING

STAVROS CHATZIPAVLIDIS

CONTENTS

Abstract

This thesis challenges the prevalent focus on achieving high accuracy in loan default prediction by addressing the inherent challenge of highly imbalanced datasets, where the minority class, particularly in loan default scenarios, holds substantial significance due to elevated costs associated with misclassifications, specifically False Negatives (FNs). To tackle this challenge, the study adopts a weighted approach, utilizing the cost-sensitive method of class weighting. Acknowledging the limitation of quantifying FN reduction by raw numerical change, a novel profit-maximizing custom scoring function is introduced. This function assigns distinct costs to True Negatives (TNs) and FNs, aiming to identify a model that maximizes overall profitability. The models investigated include Logistic Regression (LR) as the baseline, along with weighted variations of LR, Decision Tree (DT), Random Forest (RF), and state-of-the-art models CatBoost (CB) and XGBoost (XGB). Results highlight the impact of employing class weights and the custom scoring function, yielding a 71.63% to 82.48% improvement in accurately classifying the minority class. Crucially, this improvement translates to a 737%-1031% surge in profitability. Weighted XGBoost (WXGB) emerged as the most profitable model in this study, significantly outperforming the baseline. It achieved a 80% reduction in false negatives and generated a substantial profit of $45,595,661. In contrast, the baseline model resulted in a negative profit of -$4,897,414, showcasing a 1031% increase in profitability when compared to the baseline. Findings highlight the efficacy of class weighting and a profit-maximizing custom scoring function in enhancing profitability in scenarios with class imbalance and varying error costs.

DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

Data Source:

The data utilized in this thesis was obtained from Kaggle, with the original author credited as R.G. The data owner is identified as Dark_Raider. It is important to note that the dataset is publicly available under the GPL 2 license, and the author adheres to the terms and conditions outlined by the license.

Ethics:

This thesis project does not involve the collection of data from human participants or animals. The data used is sourced from a publicly available dataset on Kaggle, and ethical considerations are maintained by adhering to the terms of use specified by the data owner and the GPL 2 license.

Code:

In this thesis, figures and graphs have been created by the author for the purpose of visualizing and presenting data obtained from cited sources.[1] The code implementation is conducted with transparency, and the thesis includes a comprehensive list, including version numbers, of all libraries and frameworks employed in the analysis.

Technology:

No external tools or services were utilized in the creation, paraphrasing, spelling or grammar checking, typesetting, or reference management of the provided text. The responses were generated based on the author's understanding and knowledge without relying on additional tools or resources.

1 INTRODUCTION

As stated by the Basel Committee on Banking Supervision (2000), default risk, in its most basic form, refers to the likelihood that a borrower or other entity associated with a bank will be unable to meet its financial commitments in line with the mutually determined terms. The core purpose of default risk management is to maximize a bank's profitability by keeping its vulnerability to uncertainty about obligation fulfillment well below the required level.

1.1 *Peer-to-peer Lending*

According to Bachmann et al. (2011), peer-to-peer (P2P) lending is a contemporary method of obtaining loans that circumvent conventional financial institutions. It operates through online platforms where individuals serve

[1] All original data sources are appropriately cited and referenced in the bibliography. The figures and graphs presented herein are the author's original work for the purpose of illustration.

as both lenders and borrowers, engaging in direct lending transactions. These platforms facilitate connections between individuals seeking loans and those willing to invest, creating a dual-sided market. Borrowers provide information about their loan requests and financial standing, while lenders propose loans with interest rates determined by this information. The P2P model enables borrowers to secure loans without the involvement of traditional financial intermediaries, potentially leading to more favorable terms. As per Lenz (2016), the platform strategically navigates away from the conventional methods employed by commercial banks, skillfully avoiding the assumption of credit risk in its contractual positions. Unlike the centralized approach of banks, which amalgamate credit and liquidity risks in loans funded by deposits and liabilities, the platform opts for a decentralized strategy, dispersing these risks among its user base.

In contrast to banks, the platform does not face direct exposure to credit risk in its contractual dealings. The conventional banking model heavily relies on the interest margin between deposit and loan rates. In contrast, the platform's business model remains resilient to market interest rate fluctuations, generating revenue through transaction fees imposed on both borrowers and lenders. This distinctive model underscores the platform's unique risk distribution strategy.

1.2 *Rising Debt and Delinquency Rates*

Figure 1 illustrates that the household debt skyrocketed from 13.22 trillion dollars in Q1 2018 to a record-breaking 17.06 trillion dollars. This rapid surge underscores the importance of monitoring and managing household debt, as it can significantly impact financial stability, economic health, and the well-being of individuals and families throughout the United States.

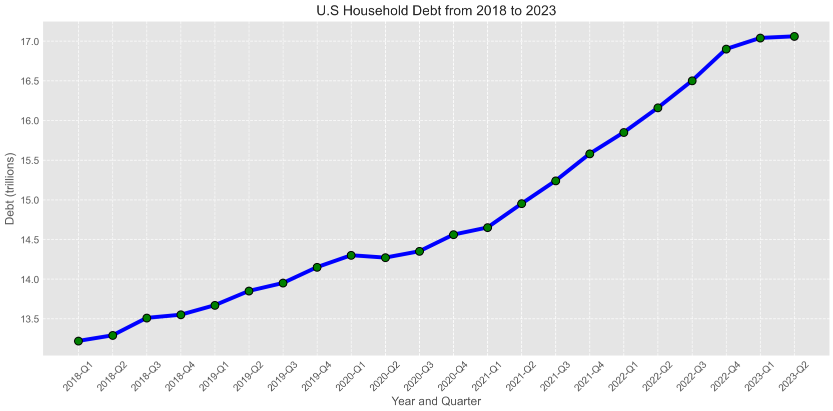


Figure 1: U.S Household Debt from 2018 to 2023 (Federal Reserve Bank of New York Research and Statistics Group, 2023)

The increasing household debt emphasizes the need for more efficient predictive mechanisms to identify potential delinquents. Table 1 shows that serious delinquencies rose by 38.10% between 2022 and 2023. In this context, the delinquency rate is an essential indicator, demonstrating borrowers' capacity to fulfill their financial commitments and emphasizing the need for robust tools and models to accurately anticipate and identify individuals at risk of default. Improving predictive mechanisms is critical in mitigating the impact of high debt levels and encouraging better risk management strategies.

Debt Category	Q2 2022	Q2 2023
Mortgage Debt	0.44%	0.63%
Home Equity Line Of Credit	0.32%	0.44%
Student Loan Debt	0.98%	0.85%
Auto Loan Debt	1.81%	2.41%
Credit Card Debt	3.35%	5.08%
Other	3.21%	4.65%
ALL	0.84%	1.16%

Table 1: Serious Delinquency Rates (90 days or more delinquent) in Q2 2022 and Q2 2023 (Federal Reserve Bank of New York Research and Statistics Group, 2023)

1.3 *Early Machine Learning Adoption in Loan Default*

Credit risk analysis, a fundamental component of financial decision-making, has experienced a notable evolution with the integration of Machine Learning (ML) techniques. Conventional creditworthiness evaluation approaches, which depend on linear computations of a restricted array of metrics, have demonstrated inconsistencies and a lack of reliability in their outcomes (Ereiz, 2019). A previously published article (Langley & Simon, 1995) reported that a financial service corporation, in managing loan applications, had employed a data-driven analysis approach grounded in quantitative decision-making methodologies. This method entailed refusing candidates who fell below an established limit and accepting those who exceeded another. Approximately 10% to 15% of the applicants were classified as grey area cases, requiring loan supervisors to evaluate them before making a final decision. Surprisingly, reports revealed that loan supervisors were barely fifty percent accurate when forecasting whether or not these grey-area applicants would fail to repay their loans. The results of this research drove this financial institution to investigate ML techniques to enhance the decision-making process.

In the initial phases of default risk assessment, LR emerged as a prevailing method (Gray, 1985), representing the techniques of that era. Within the domain of credit assessment, ensemble techniques rooted in decision trees, such as the RF approach, outperform conventional LR models in classification efficacy.

Nevertheless, LR is the industry's gold standard for assessing credit risk. This enduring preference primarily arises from the inherent obscurity of ensemble methods, which fail to align with the stringent interpretability demands set by financial oversight bodies (Dumitrescu et al., 2021).

Given the critical advantages of LR, such as its exceptional interpretability, historical role as an industry standard, and model simplicity, it is a natural selection for our research as the baseline model in credit risk assessment. A solid foundation will be laid by using LR as the baseline, which prioritizes model interpretability and simplicity while also serving as a benchmark against which more complex models can be compared for predictive accuracy.

As ML models gain popularity, the need for model interpretability becomes even more pronounced due to the increasing complexity of these models, which poses a challenge to understanding their decision-making processes. To address this issue, researchers, as demonstrated by Gramegna and Giudici (2021), have employed cutting-edge techniques such as SHapley Additive exPlanations (SHAP) values and Local Interpretable Model-agnostic Explanations (LIME). These techniques enable the use of advanced ML models with improved performance without the loss of interpretability. These methodologies not only allow the use of advanced ML models with improved performance but also significantly enhance transparency and accountability.

1.4 *Challenges in Credit Risk Assessment*

One of the biggest challenges in addressing loan default is class imbalance in scenarios where one class significantly outnumbers others; a condition known as class imbalance arises. The prevalent class is often called the majority class, while the less frequent class is designated the minority class. However, in various applications, the instances of the minority class hold more significance and relevance. The imbalance issue becomes particularly pronounced when the class of interest is rare, with fewer occurrences than the majority class (Guo et al., 2008). Additionally, the consequences of misclassifying the minority class carry a substantially higher cost than misclassifying the majority class, as seen in examples like cancer versus non-cancer or fraud versus non-fraud (Naganjaneyulu & Kuppa, 2013). Several research investigations have indicated that constructing models on imbalanced datasets may lead to elevated specificity or local accuracy for the majority class while yielding suboptimal outcomes for the minority class across the same metrics (Fernandez, Garcia, Herrera, & Chawla, 2018). To resolve the class imbalance challenge, researchers have often used the cost-sensitive learning technique of class weighting (Phan & Yamamoto, 2020; L. Zhu, 2019; M. Zhu et al., 2018).

1.5 *Motivation and Goal*

Predicting loan defaults accurately, rather than misclassifying them, has far-reaching implications. Accurate predictions improve financial systems and risk management tools, optimizing decision-making. They also contribute to overall financial market stability by ensuring responsible lending practices and economic fairness on a societal level. While there is a notable focus on accuracy in predicting loan defaults during the loan approval process, it is equally crucial to address the oversight of potential defaulters (Aslam et al., 2019). This oversight presents significant financial and societal challenges.

This study aims to minimize FNs in loan assessments, promoting financial stability, integrity, and inclusivity. Lenders' economic interests are protected by reducing instances where loans are mistakenly classified as non-defaults, and borrowers facing genuine financial difficulties receive the necessary attention. Additionally, this effort contributes to the more equitable distribution of financial resources, preventing economic disparities and fostering a resilient and just financial ecosystem. The goal is to elevate the conversation about responsible lending practices and contribute to creating a financial framework that genuinely serves all.

1.6 *Research Strategy*

1.6.1 *Research Questions*

Main Research Question

How does integrating class weights to address the class imbalance, along with a profit-maximizing custom scoring function, influence the effectiveness of ML models in minimizing false negatives in the context of loan default prediction?

Sub-Research Questions

1. *To what extent does including the custom scoring function and class weights impact the predictive performance of ML models?*

2. *Which machine learning model, among those considered, achieves the highest profitability in predicting loan default with class weights and the custom scoring function, emphasizing real-world financial implications?*

1.6.2 *Research Findings*

The investigation into the impact of incorporating class weights and a custom scoring function on machine learning models revealed significant improvements in predictive performance. Including these elements led to noticeable improvements in G-mean, macro-average F1, class 1 recall, and AUC scores across all models, signaling an overall enhancement in their predictive capabilities. Notably, Weighted Random Forest (WRF), Weighted CatBoost (WCB), and WXGB stood out for their heightened performance in correctly classifying the minority class 1, compared to the LR baseline, which consistently misclassified instances of the minority class. However, this boost in predictive performance came with a trade-off, resulting in an increased misclassification rate for class 0 instances. Despite this trade-off, the top-performing models, WCB and WXGB, demonstrated an optimal balance by accurately classifying the minority class while minimizing misclassifications of the majority class. This equilibrium is further emphasized by the profitability analysis, revealing an increase in profit ranging from 737% to 1031% across all models compared to the LR baseline.

2 RELATED WORK

2.1 *Machine Learning Models in Credit Risk Assessment*

A diverse array of ML models has been explored within the realm of credit risk assessment, each aiming to predict the probability of loan default. Notable investigations include:

- the utilization of Support Vector Machines (SVM) (Moula et al., 2017);
- the application of Neural Networks (NN) (Bayraci & Susuz, 2019);
- the employment of K-Nearest Neighbors (Mukid et al., 2018);
- the exploration of Naive Bayes (NB) methodologies (Krichene, 2017); and
- the investigation of tree-based models (Madaan et al., 2021).

Additionally, within the category of tree-based models, a subcategory of gradient-boosted trees has explicitly been investigated (Ponsam et al., 2021).

This study will predominantly leverage the capabilities of tree-based models, explicitly exploring the performance of boosted-tree ensemble models, including XGB and CB. This focus is driven by the natural resilience of tree-based models to outliers present in predictor variables (Sigrist & Hirnschall, 2019). Moreover, when compared with deep learning, tree-based models excel at classification tasks. This superiority can be attributed to the unique characteristics of tabular data, such as irregular patterns in the target function, uninformative features, and non-rotationally invariant data (Grinsztajn et al., 2022).

In the domain of loan default prediction, L. Zhu et al. (2019) explored the predictive capabilities of diverse models, including SVM, DT, and RF, for loan default prediction, with RF emerging as the superior performer.

Furthermore, Ibrahim et al. (2020) conducted a comprehensive study on loan approval prediction, evaluating models such as LR, RF, Adaboost, XGB, NN, Gradient Boosting (GB), CB, and DT. Their findings highlighted CB as the top-performing model, showcasing the highest F1 score, Area Under the Curve (AUC), and precision in loan approval prediction. In a study by Barbaglia et al. (2023) focused on loan default forecasting, various predictive models, including XGB, GB, RF, Penalized Logistic with Power Series, Penalized Logistic, NN, and NB, were assessed. The evaluation results identified XGB as the model with superior performance in loan default forecasting, achieving notably outstanding forecasting accuracy on out-of-sample data. A recurring theme in the literature review is the consistent outperformance of tree-based models, particularly boosted algorithms, over SVM, NN, and NB. The prevailing tendency suggests that XGB and CB generally demonstrate top-tier performance, often with closely matched results.

2.2 *Differential Impact of Prediction Errors*

Understanding the consequences of prediction errors is paramount in various fields, ranging from healthcare (Naganjaneyulu & Kuppa, 2013) to finance (Harvey & Liu, 2020) and beyond. This section explores the differential impact of different types of errors, emphasizing the significance of TNs and FNs in the loan default context.

Errors and Costs in Loan Default Prediction

The research by N. Chen et al. (2016) highlights the presence of cost sensitivity in credit risk assessment, signifying a discrepancy in misclassification costs. Specifically, the repercussions of erroneously categorizing a credit as positive (indicating good credit) when it is negative (indicating bad credit) typically outweigh the costs associated with misclassifying good credit as bad. To address the challenge of class imbalance, the study recommends the utilization of a cost-sensitive classifier. In this approach, the misclassification cost for the minority class is deliberately set higher than that for the majority class during training. Other researchers have also underscored the significance of considering the costs of FNs and TNs. They advocate for assigning a higher weight to instances where loans are predicted as fully paid but, in reality, are defaulted (Cao et al., 2013; L. Zhu, 2019).

2.3 *Imbalanced Data Classification*

The prevalence of imbalanced data spans diverse domains, including the medical field (Rahman & Davis, 2013), credit card fraud detection (Priscilla & Prabha, 2020), and cybersecurity (Leevy et al., 2021). Notably, the domain of loan defaults stands out as a significant challenge characterized by class imbalance, as evidenced by several studies (Y. Chen et al., 2021; Coser et al., 2019; Namvar et al., 2018; Shingi, 2020; L. Zhu, 2019). Addressing this challenge has prompted an increase in research dedicated to the classification of imbalanced data (Ali, Salleh, Saedudin, et al., 2019; Johnson & Khoshgoftaar, 2019; Rezvani & Wang, 2023). As outlined by Ali, Salleh, Saedudin, et al. (2019), addressing imbalanced data involves employing four critical solutions:

- preprocessing techniques (Ali, Salleh, Hussain, et al., 2019);
- algorithmic approaches (Fernandez, Garcia, Galar, et al., 2018);
- cost sensitivity (Liu & Zhou, 2006); and
- ensemble learning (Feng et al., 2018).

In this study, the cost-sensitive learning technique of class weights will be employed for this purpose. Referencing a study by M. Zhu et al. (2018), integrating class weights with the RF model is explored to address imbalances in a medical dataset. Class weighting demonstrated improved overall classifier performance, achieving high accuracy in both the majority and minority classes. Moreover, in a similar context of loan default prediction, in a study by L. Zhu (2019), techniques such as SMOTE and ADASYN, along with cost-sensitive learning approaches like class weights, were applied across LR, RF, and XGB models. Notably, the WXGB emerged as the top performer, providing strong motivation for incorporating class weights in the present study.

2.4 *Critical Gap in the Literature*

According to Aslam et al. (2019), the existing literature predominantly focuses on accurately identifying actual loan defaults, leaving a notable gap in understanding the consequences of missing such defaults, particularly regarding FNs. The recognized gap underscores the critical importance of investigating the potentially harmful effects of FNs on lending organizations, a focal point in this research. While the literature highlights the increasing effectiveness of ML techniques in credit risk analysis, a significant limitation persists—the oversight of FNs. This thesis aims to bridge this gap by recognizing the imperative need to address FNs, particularly in the context of loan default prediction.

3 METHOD

3.1 *Custom Scoring Function for Loan Default Prediction*

This approach is motivated by a study (Serrano-Cinca & Gutierrez-Nieto, 2016) that deviates from the conventional emphasis on predicting the probability of default in P2P lending. Rather than focusing on loan defaults, the study aims to forecast the expected profitability of investing in P2P loans. Notably, the study concludes that a lender using a profit scoring system through multivariate regression outperforms the conventional LR-based credit scoring system. This departure from traditional methods demonstrates the effectiveness of a profit-driven approach in improving P2P loan investment decision-making.

In this thesis, a custom scoring function tailored to the specific considerations of the P2P lending market is introduced. This scoring function aims to balance minimizing the cost associated with FNs and maximizing the profit from TNs.

In the event of borrower defaults, resulting in lenders carrying the losses, the platform, although not liable for these losses, is often obligated to manage missed payments (Jorgensen, 2018). Defaulted loans frequently require legal intervention for potential recovery, which can prove challenging and time-consuming. This study's scenario revolves around non-repaid loans (FNs), where the remaining amount is irrecoverable. This loss is partially mitigated by the interest rate collected up to the point of default.

On the other hand, TNs indicate loans predicted as non-defaulting and successfully repaid. The market profits from the interest collected, either the total amount if the loan is paid off or a portion if the loan is current.

The custom scoring function is defined as follows:

$$\text{Profit from TNs} = \begin{cases} \frac{\text{LD}}{\text{T}} \times \left(\frac{\text{IR}}{100}\right) \times \text{LA}, & \text{if actual = 0 and predicted = 0} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where LD is the current loan duration, T stands for the loan term, IR denotes the interest rate, and LA is the loan amount.

This equation calculates the profit from TNs, considering both repaid and current loans, where the interest collected contributes to the overall profit.

$$\text{Loss from FNs} = \begin{cases} \left(\frac{\text{T}-\text{LD}}{\text{T}}\right) \times \text{LA} - \left(\frac{\text{LD}}{\text{T}} \times \frac{\text{IR}}{100} \times \text{LA}\right), & \text{if actual = 1 and predicted = 0} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

This equation computes the loss resulting from FNs. It addresses instances where a loan defaults before the specified term, incurring a loss equivalent to the remaining amount to be repaid. This loss is partially mitigated by the interest rate collected up to the point of default.

Combining Equations (1) and (2), we get:

$$\text{Custom Scoring Function} = \text{Profit from TNs} - \text{Loss from FNs} \tag{3}$$

The objective is to find a model that maximizes this custom scoring function, capturing the optimal balance for the given business scenario.

3.2 *Confusion Matrix*

The confusion matrix is a tool to evaluate the performance of a classification model. In the context of credit risk assessment:

	Predicted Non-Default	Predicted Default
Actual Non-Default	True Negative (TN)	False Positive (FP)
Actual Default	False Negative (FN)	True Positive (TP)

Table 2: Confusion Matrix

True Positive (TP): Predicted default and actually defaulted

True Negative (TN): Predicted non-default and actually non-defaulted

False Positive (FP): Predicted default but actually non-defaulted

False Negative (FN): Predicted non-default but actually defaulted

3.3 *Performance Metrics*

3.3.1 *Geometric Mean*

According to several researchers (Guo et al., 2008; Weiss & Provost, 2003), the emphasis on accuracy tends to prioritize the majority class over the minority class, posing a challenge for classifiers to handle the minority class effectively. On the other hand, the geometric mean (G-mean) is calculated by multiplying the accuracy scores for both classes: sensitivity (accuracy in positive instances) and specificity (accuracy in negative instances). This statistic implies a balance in classification efficiency between the majority and minority classes. Suppose the model exhibits poor prediction capacity for positive cases. In that case, the G-mean value decreases even if the negative examples are correctly identified, emphasizing the model's sensitivity to performance discrepancies between the two classes (Hido et al., 2009).

The G-mean is calculated using the formula:

$$\text{G-mean} = \sqrt{\text{Sensitivity} \times \text{Specificity}} \tag{4}$$

3.3.2 *Macro-Average F1*

Employing macro-average metrics allocates equal significance to each class, regardless of their frequencies, showcasing the model's resilience in confronting imbalanced data challenges (Riyanto et al., 2023). Specifically, macro-average F1 computes the F1 score for each class individually, offering a comprehensive assessment of the model's effectiveness across all classes and providing a more balanced evaluation of its performance.

Let P_i and R_i represent precision and recall for class i, and N be the number of classes.

The F1 score for class i is given by:

$$F1_i = 2 \times \frac{P_i \times R_i}{P_i + R_i} \tag{5}$$

$$Macro\_F1 = \frac{1}{N} \sum_{i=1}^{N} F1_i \tag{6}$$

3.3.3 *Recall per Class*

In highly imbalanced class distributions, the recall for the minority class is frequently observed to be zero (Guo et al., 2008). This phenomenon contributes to an inflated overall recall rate, predominantly influenced by class imbalance. Consequently, this thesis will prioritize the recall assessment for class 1, with an ongoing exploration of recall for class 0.

The objective is to investigate how enhancements in predictive performance for class 1 may impact the predictions for class 0. Recall for each class is calculated as follows:

$$\text{Recall Class 0} = \frac{\text{True Negatives Class 0}}{\text{True Negatives Class 0} + \text{False Positives Class 0}} \tag{7}$$

$$\text{Recall Class 1} = \frac{\text{True Positives Class 1}}{\text{True Positives Class 1} + \text{False Negatives Class 1}} \tag{8}$$

3.4 *Class Weights*

Class weights are assigned to each class based on their frequency in the training data. The imbalance parameter a is used to adjust these weights.

Imbalance Parameter (a)

The imbalance parameter a allows fine-tuning the influence of class weights. A higher a value increases the emphasis on the minority class during training, aiding the model in capturing patterns within the less-represented class.

Class Weight Calculation

The class weight for each class is calculated as follows:

$$\text{Class Weight for Class } i = \frac{m}{a \cdot n_i} \tag{9}$$

where m is the total number of samples, a is the imbalance parameter, and n_i is the number of samples in class i.

The study will explore the specific formulation of class weights in different ML frameworks in later sections. At its core, the overarching concept involves adjusting the contribution of each class to the overall loss function during the training phase.

3.5 *K-Fold Cross-Validation*

In K-fold cross-validation, the dataset is initially divided into K distinct subsets or blocks, all of equal size (Bengio & Grandvalet, 2003). One set among the subsets functions as the validation set, while the remaining subsets form the training set used for model training. Specifically, concerning the kth subset, the model is trained on the rest of the data (K - 1 subsets), and the prediction error of the fitted model is computed when predicting within this kth subset. This iterative process spans $k = 1, 2, \ldots, K$, concluding in the aggregation of K estimates of prediction error (Fushiki, 2011). Figure 2 illustrates a single iteration of K-fold cross-validation with multiple folds. In this study, given the imbalanced nature of the dataset, The choice was made to employ stratified k-fold cross-validation, ensuring consistent class imbalance ratios in each fold. This approach promotes a robust representation of every class throughout the validation process (Ariza-Garzon et al., 2020).

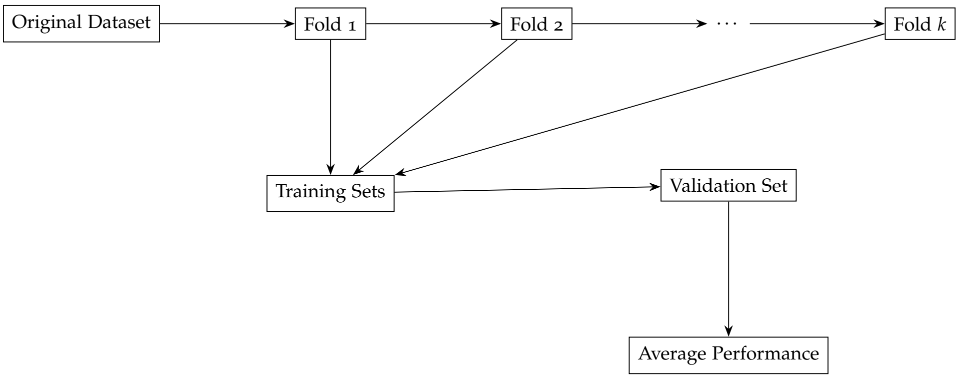


Figure 2: Single Iteration of K-Fold Cross-Validation with Multiple Folds
Note: Created by the author with methodology as described by Fushiki (2011)

3.6 *Logistic Regression*

This thesis assumes a foundational understanding of LR, DT, and RF algorithms, progressing to explore their customizations.

3.6.1 *Binary Cross-Entropy*

The LR model is trained by minimizing the Binary Cross-Entropy (BCE) loss function, defined as:

$$\text{BCE} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \tag{10}$$

where N is the number of samples, y_i is the true label for sample i, and p_i is the predicted probability of default for sample i.

3.6.2 *Weighted Binary Cross-Entropy*

To include a weighted version of the binary cross-entropy, according to Saber et al. (2022), the equation becomes:

$$WBCE = -\sum_{i=1}^{m} \left(ay_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)\right) \tag{11}$$

The cost function `WBCE` uses a as the imbalance parameter to adjust class weights. In this context, m denotes the number of samples in the dataset, y_i represents the true label of the i-th sample (0 or 1), and $\hat{y}_i$ signifies the predicted probability that sample i belongs to class 1.

3.6.3 *Hyperparameter Exploration: Weighted Logistic Regression*

*Note: In each future hyperparameter exploration table, the best parameter values for each model will be highlighted with **bold text**.*

Table 3: Hyperparameter Exploration: Weighted Logistic Regression

Hyperparameter	Description	Domain
`class_weight`	Class imbalance handling	balanced, {0: 1, 1: 6, 8, 10, **12**, 14}
`C`	Regularization strength	0.001, **0.01**, 0.1, 1
`penalty`	Type of regularization	’l1’, **’l2’**
`solver`	Logistic regression solver	**’liblinear’**

3.7 *Decision Tree*

3.7.1 *Weighted Entropy*

According to Park et al. (2017), for a set of clusters $C_0, \ldots, C_{N-1}$, the weighted entropy S is defined as:

$$S = -\sum_{n} I_n \log P_n \tag{12}$$

where P_n represents the relative frequency of weights in the range of values for cluster C_n, calculated as the ratio of the number of weights in C_n to the total number of weights in all clusters:

$$P_n = \frac{|C_n|}{\sum_k |C_k|} \tag{13}$$

Moreover, I_n is the representative importance, denoting the average importance of all weights in cluster C_n :

$$I_n = \frac{\sum_m i(n, m)}{|C_n|} \tag{14}$$

where $i(n, m)$ is the importance of weight m in cluster C_n.

3.7.2 *Weighted Gini Impurity*

Gini impurity assesses how effectively a split divides binary-class samples within a specific node. According to Disha and Waheed (2022) it is expressed as:

$$i(\tau) = 1 - P_p^2 - P_n^2 \tag{15}$$

where P_p is the fraction of positive samples, and P_n is the fraction of negative samples out of the total (N) samples at node τ.

The weighted Gini impurity for each class independently at node τ is given by:

$$i_{\text{weighted}}(\tau) = 1 - \sum_c w_c \cdot P_c^2 \tag{16}$$

where $i_{\text{weighted}}(\tau)$ represent the weighted gini impurity at node τ, where w_c denotes the weight associated with class c, and P_c represents the fraction of samples belonging to class c at node τ.

3.7.3 *Hyperparameter Exploration: Decision Tree*

Table 4: Hyperparameter Exploration: Decision Tree

Hyperparameter	Description	Domain
`criterion`	Splitting criterion	**'gini'**, 'entropy'
`max_depth`	Maximum depth of the tree	**20**, 40
`min_samples_split`	Minimum number of samples required to split an internal node	2, **5**
`min_samples_leaf`	Minimum number of samples required to be at a leaf node	**1**, 2
`class_weight`	Class imbalance handling	**balanced**, {0: 1, 1: 6, 8, 10, 12, 14}

3.8 *Random Forest*

3.8.1 *Aggregated Weighted Entropy*

Incorporating Equation 12 as discussed in sub-subsection 3.7.1, the aggregated weighted entropy for the WRF can be expressed as follows:

$$S_{\text{avg}} = \frac{1}{T} \sum_{t=1}^{T} S_t \tag{17}$$

Where T represents the number of trees in the forest.

3.8.2 *Aggregated Weighted Gini Impurity*

According to Disha and Waheed (2022) the reduction in Gini impurity resulting from an optimal split $\Delta i_f(\tau, M)$ is aggregated across all nodes (τ) in M weighted trees in the forest for each feature:

$$Ig(f) = \sum_{M} W_{p,n} \sum_{\tau} \Delta i_f(\tau, M) \tag{18}$$

Here, Ig represents Gini importance, indicating how often a feature f is chosen for a split and its overall discriminative significance for binary classification. The weight $W_{p,n}$ defines an imbalanced class distribution in the learning algorithm.

Weight adjustments for positive and negative classes are defined as:

$$\text{Weight for positive class: } W_p = \frac{n_n}{N} \tag{19}$$

$$\text{Weight for negative class: } W_n = \frac{n_p}{N} \tag{20}$$

Where n_n is the number of negative instances, n_p is the number of positive instances, and N is the total number of instances in the training dataset. The weights satisfy $W_p + W_n = 1$, and in the case of imbalanced class data, $W_p \neq W_n$.

3.8.3 *Hyperparameter Exploration: Random Forest*

Table 5: Hyperparameter Exploration: Random Forest

Hyperparameter	**Description**	**Domain**
n_estimators	Number of trees in the forest	**400**, 500
criterion	Function to measure the quality of a split	**’entropy’**, ’gini’
max_depth	Maximum tree depth	**20**, 40
min_samples_split	Minimum number of samples required to split an internal node	2, **5**
min_samples_leaf	Minimum number of samples required to be at a leaf node	1, **2**
class_weight	Class imbalance handling	**balanced**, {0: 1, 1: 6, 8, 10, 12, 14}

3.9 *CatBoost*

3.9.1 *Model Introduction*

CB, a gradient-boosting ML library introduced by Dorogush et al. (2017), uses oblivious decision trees to ensure balance, reduce overfitting, and accelerate testing execution by maintaining the same splitting criterion across an entire tree level. It employs an iterative strategy for tree construction, progressively combining categorical features. Numeric values for these combinations are dynamically generated in real-time, extending to merging numerical and categorical features and treating specific splits as categorical. It also introduces an alternative technique for transforming categories into numerical values by assessing the frequency of each category in the dataset—a method similarly applied to feature combinations (Dorogush et al., 2017). Efficiently managing categorical features during tree construction, CB utilizes an ordered boosting scheme that incorporates strategies such as random permutation and the calculation of average label values, leading to a reduction in overfitting and an overall enhancement in algorithmic performance (Prokhorenkova et al., 2018). According to Dorogush et al. (2017), CB diverges from XGB in managing gradient bias during tree construction. Unlike XGB, where tree structure and leaf values are determined sequentially, CB divides this process. Initially, it utilizes a modified strategy to select the tree structure and address gradient bias. The subsequent phase follows a conventional approach in setting leaf values. This distinctive method in CB aims to present a more effective solution for countering gradient bias than XGB.

CB is driven by minimizing the designated function $\mathcal{L}$ throughout the training process. This objective function serves as a gauge for the error or deviation between predicted values and true labels. Specifically tailored for binary classification tasks, the default objective is binary cross-entropy—an essential metric that encapsulates the negative log-likelihood of true labels considering the predicted probabilities. These gradient-boosting frameworks aim to craft an ensemble of trees that collectively minimize binary cross-entropy, enhancing predictive performance for binary classification scenarios. The expressions for BCE and WBCE are denoted by Equation 10 and Equation 11 in sub-subsections 3.6.1 and 3.6.2, respectively.

3.9.2 *Hyperparameter Exploration: CatBoost*

Table 6: Hyperparameter Exploration: CatBoost

Hyperparameter	Description	Domain
`depth`	Tree depth	6, 8, **10**, 12
`learning_rate`	Learning rate for boosting	0.05, **0.1**, 0.15
`l2_leaf_reg`	L2 regularization coefficient for leaf values	**3**, 5, 7
`iterations`	Boosting iterations	200, **300**, 400, 500
`class_weights`	Class imbalance handling	**balanced**, {0: 1, 1: 6, 8, 10, 12, 14}
`bootstrap_type`	Bootstrap type	Bayesian, **Bernoulli**, MVS

3.10 *XGBoost*

3.10.1 *Model Introduction*

XGB, or eXtreme Gradient Boosting, introduced by T. Chen and Guestrin (2016), applies a set of classification and regression trees as weak learners and then boosts the performance of the trees by creating an ensemble of trees that minimizes a regularized objective function (Fauzan & Murfi, 2018). XGB employs a process known as pre-sorting; categorical features are efficiently encoded before constructing decision trees (Alshari et al., 2021). According to T. Chen and Guestrin (2016), it employs advanced node splitting and depth control techniques in its tree construction process. In constructing trees, XGB optimizes distributed computing through parallelization, adopts a column block structure for efficient memory usage, and employs a sophisticated weighted quantile sketch algorithm to precisely identify optimal split points (T. Chen & Guestrin, 2016). The algorithm's learning objective incorporates regularization, balancing training loss, and penalty terms to prevent overfitting and enhance model generalization. Notably, the node-splitting algorithm intelligently determines optimal splits, utilizing a unique technique that considers second-order gradient information, ensuring a comprehensive exploration of the feature space. Prediction involves summing contributions from individual trees, with shrinkage applied to each tree's output to prevent overfitting (T. Chen & Guestrin, 2016). The application of these sophisticated techniques, as outlined in the paper (T. Chen & Guestrin, 2016), establishes XGB as a potent tool for predictive modeling, particularly in scenarios involving large datasets and intricate relationships.

Similar to CB, XGB utilizes the BCE and WBCE. Refer to Equations 10 and 11 in sub-subsections 3.6.1 and 3.6.2 for the respective loss function expressions.

3.10.2 *Hyperparameter Exploration: XGBoost*

Table 7: Hyperparameter Exploration: XGBoost

Hyperparameter	Description	Domain
`max_depth`	Maximum depth of a tree	6, 8, 10, **12**
`learning_rate`	Learning rate for boosting	0.05, **0.1**, 0.15
`reg_lambda`	L2 regularization term	**3**, 5, 7
`n_estimators`	Boosting iterations	**200**, 300, 400, 500
`scale_pos_weight`	Class imbalance handling	balanced, {0: 1, 1: 6, 8, 10, **12**, 14}
`colsample_bytree`	Fraction of features to be randomly sampled for each tree	**0.7**, 0.8, 0.9, 1
`reg_alpha`	L1 regularization term	0.1, 0.2, **0.3**

3.11 *Feature Contribution Analysis using SHAP*

In interpreting ML predictions, additive feature attribution methods represent a prevalent category, explaining a model's output as a sum of actual values for each input feature (Lundberg et al., 2018). Specifically, these methods use an explanation model g expressed as a linear function of binary variables:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \tag{21}$$

Equation 21 can interpret an individual prediction and the entire model by considering the average feature attribution across all observations (Meng et al., 2020).

Motivating the use of SHAP values for tree ensemble feature attribution, the method leverages conditional expectations ($E[f(x)|xS]$), where S denotes the set of non-zero indexes in z' (Lundberg et al., 2018). SHAP values combine these conditional expectations with classical Shapley values from game theory to attribute values (ϕ_i) to each feature:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} [f(S \cup \{i\}) - f(S)] \tag{22}$$

In tree models, SHAP values emerge as the sole reliable and locally precise method that adheres to the missingness property and employs conditional dependence to measure missingness (Lundberg et al., 2018).

4 EXPERIMENTAL SETUP

4.1 *Data description*

The dataset used in this study, the' Credit Risk Analysis' dataset, is sourced from Kaggle and publicly available under the GPL 2 license (R.G., 2021). It comprises 887,379 instances and 74 columns from the well-known peer-to-peer lending website LendingClub. This large dataset is accessible in CSV format and contains various loan information, including loans granted for various purposes. It contains essential attributes such as loan length, loan amount, loan grade, interest rate, house ownership status, annual income, loan purpose, and loan status—the latter of which acts as the target label for predictive modeling. The dataset also includes payment plan indicators and delinquency-related data. It covers the period from December 31, 2006, to December 30, 2015.

4.1.1 *Data preproccessing*

Through feature engineering, 19 features were selectively retained from an initial pool of 74. This selection was based on careful considerations such as missing values and the degree of correlation. Features without critical relevance to the analysis and those exhibiting substantial missing values and high correlations were removed. Categorical variables were transformed into dummy variables, and `grade` was encoded using label encoding. A new variable, `loan_duration_months`, was introduced based on loan initiation dates for temporal analysis. Missing values were imputed with feature medians. Outliers were removed for specific variables (`revol_util`, `annual_income`, `dti`, `revol_bal`) to enhance dataset distribution, focusing on the most extreme values likely associated with data entry errors. Only 77 instances of outliers were deleted out of 887,379 instances, contributing to a cleaner and more reliable dataset for analysis. Tables 13 and 14 in Appendix A (page 41) provide a detailed summary of the dataset's predictors and its missing values.

4.1.2 *Exploratory Data Analysis*

The Exploratory Data Analysis (EDA) provides a comprehensive dataset overview, highlighting key aspects such as outliers, correlations, and trends. The loan distribution indicates a highly imbalanced structure, with non-defaulted loans constituting 91.2% and defaulted loans forming only 8.8%.

The numerical Pearson correlation matrix reveals a significant correlation of 0.94 between `installment` and `loan_amnt`, suggesting potential multicollinearity. Nominal predictor variables, analyzed using Cramér's V correlation coefficients, demonstrate minimal correlation, ranging from 0 to 0.11, indicating low associations and relative independence.

Additionally, the categorical predictor `grade` exhibits a noticeable trend, showing a progressively higher average interest rate correlated with loan grade. This observation underscores the ordinal nature of the variable and provides insights into the ascending pattern of interest rates based on loan grades.

The visuals, including class distribution, the resulting outlier box plot, Pearson correlation matrix for numerical predictors, Cramér's V correlation matrix for nominal predictors, and the interest rate trend across loan grades, can be referenced in Figures 8, 9, 10, 11, 12 in Appendix B (page 42).

4.2 *Data Partitioning and Validation Approach*

The final dataset splitting allocated 80% of the data to the training set, 10% to the validation set, and another 10% to the test set. This deliberate distribution aims to provide ample data for model training while retaining separate sets for fine-tuning (validation) and unbiased evaluation (test).

In this study, a 5-fold stratified k-fold cross-validation is used. Including stratified K-fold cross-validation, as highlighted in a survey by Prusty et al. (2022), significantly strengthens the model by carefully handling differences in class frequencies. This strategy ensures a more reliable evaluation of the model's ability to work well with new data, as it keeps a consistent distribution of observations across various labels in each fold. Using stratified K-fold cross-validation effectively deals with class imbalances and improves the accuracy of the model's performance assessment.

4.3 *Software and Library Versions*

During my thesis, I employed Python as the programming language of choice. All coding activities were conducted exclusively in the Google Colab environment. For version control, Git 2.43.0 (Chacon & Straub, 2014) was utilized. For a comprehensive overview of the software and library versions used, please refer to Table 8.

Table 8: Software and Library Versions

Library	Version
Python (Van Rossum & Drake Jr, 1995)	3.10.12
Pandas (pandas development team, 2020)	1.5.3
NumPy (Harris et al., 2020)	1.23.5
Scikit-learn (Pedregosa et al., 2011)	1.2.2
XGBoost (T. Chen & Guestrin, 2016)	2.0.2
CatBoost (Dorogush et al., 2017)	1.2.2
Imbalanced-learn (Lemaitre et al., 2017)	0.10.1
Matplotlib (Hunter, 2007)	3.7.1
Seaborn (Waskom, 2021)	0.12.2
itertools (Van Rossum & Drake Jr, 1995)	3.10.12
shap (Schlomer et al., 2018)	0.43.0

5 RESULTS

In the results section, this study's primary focus is refining loan default prediction to maximize profit, specifically emphasizing minimizing FNs. The analysis thoroughly examines various performance metrics across multiple machine learning models, including G-mean, recall 0, recall 1, macro-average recall, ROC AUC, and F1-score. The objective is to identify the model that effectively reduces the risk of FNs in the context of loan default instances while maintaining a robust overall performance. The results conclude by pinpointing the most profitable model for loan default prediction, highlighting the economic advantages of reducing FNs in this financial context.

5.1 *Overview of Results*

The study evaluates the impact of class weights and a custom scoring function on binary classification models, including LR, Weighted Logistic Regression (WLR), Weighted Decision Tree (WDT), WRF, WCB, and WXGB. Key findings highlight that WLR significantly improves LR's ability to predict minority class instances, while WDT achieves a balanced performance in predicting both classes. WRF excels in minimizing FNs for the minority class but at the expense of a higher misclassification rate for the majority class. WCB and WXGB demonstrate noteworthy performance, consistently achieving the most balanced prediction for both classes. Incorporating class weights and a custom scoring function mitigates class imbalances and enhances model generalization.

5.2 *Performance Metrics*

5.2.1 *Confusion Matrices*

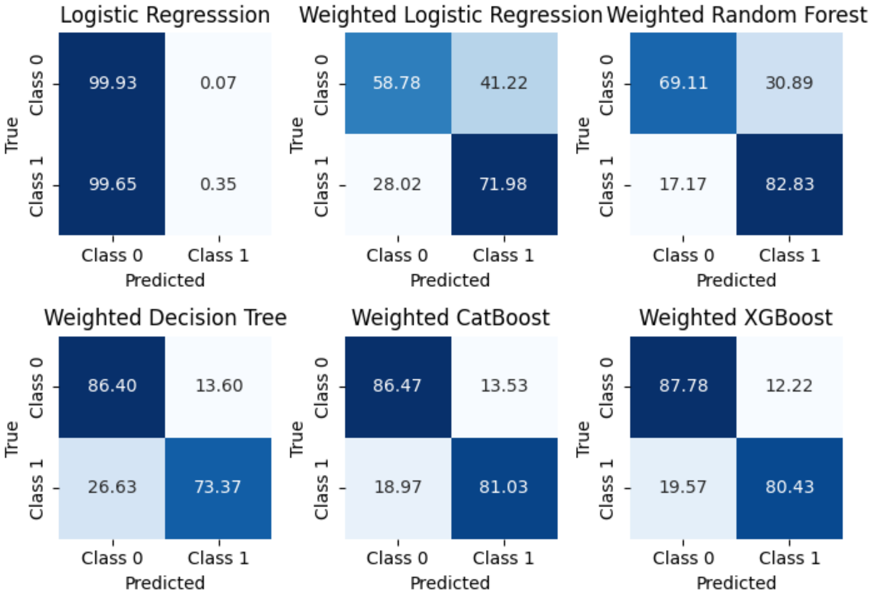


Figure 3: Performance Evaluation of Selected Models on the Test Set: Confusion Matrices with Percentage by Class

Illustrated in Figure 3, the confusion matrices of the selected models reveal insights into their performance by class percentages. LR excels in predicting class 0 with exceptional accuracy but faces challenges in class 1 predictions, misclassifying nearly all instances. WLR mitigates this imbalance by significantly decreasing FNs for class 1, even though this improvement came at the expense of a lower recall for class.

WRF excels in significantly reducing the FN rate for class 1, although it undergoes a moderate decline in accurately predicting class 0 compared to LR. WDT achieves a more balanced prediction between class 1 and class 0 compared to the previously mentioned models, demonstrating a more even distribution in minimizing FNs for class 1 while handling misclassifications for class 0.

WCB and WXGB demonstrate an ever more notable equilibrium than WDT, showing considerable reductions in FN rates for class 1 and maintaining substantial accuracy for class 0. They prioritize accurate predictions for both classes, showcasing balanced performance without overemphasizing one over the other.

For a more detailed breakdown, including the FN count per model, the percentage decrease in TNs, and the percentage decrease in FNs compared to the LR baseline model, refer to Table 9.

Model	**TN Decrease (%)**	**FN Decrease (%)**	**FNs**
Baseline Logistic Regression	-	-	7760
Weighted Logistic Regression	41.15	71.63	2182
Weighted Random Forest	30.82	82.48	1337
Weighted Decision Tree	13.53	73.02	2074
Weighted CatBoost	13.46	80.68	1477
Weighted XGBoost	12.15	80.08	1524

Table 9: Percentage decrease in True Negatives (TN) and False Negatives (FN) compared to the Logistic Regression baseline model, the number of False Negatives (FNs) is also provided

5.2.2 *Model-Specific Metrics*

In analyzing various machine learning models, LR demonstrated perfect recall for the majority class but struggled to identify instances from the minority class. This shortcoming resulted in a low geometric mean of 0.05 and a macro-average F1 score of 0.48, indicating an imbalance in its predictive capabilities. WLR showcased advancements by improving recall for the minority class, even though this improvement came at the expense of a lower recall for the majority class. This trade-off led to an enhanced geometric mean of 0.64 and a sustained macro-average F1 score of 0.48. Moreover, WDT outperformed WLR by achieving a higher recall for the majority class and maintaining a commendable recall for the minority class. The increase in correct predictions for the minority class, coupled with a high G-Mean of 0.79 and macro-average F1 score of 0.69, highlighted WDT's ability to balance both classes. Moreover, while WRF demonstrated proficiency in capturing instances from the minority class, it encountered difficulties accurately classifying instances from the majority class. This trade-off affected its overall model performance, reflected in its 0.57 macro-average F1 and 0.75 G-mean—both lower than WDT's. Notably, WCB and WXGB consistently outperformed other models, demonstrating high recall values for both classes, leading to superior geometric means (WCB: 0.83, WXGB: 0.84), macro-average F1 scores (0.71 and 0.72, respectively), and AUC values (WCB: 0.922, WXGB: 0.924). The observed consistency in performance metrics across validation and test sets suggests that overfitting was successfully mitigated, affirming the models' maintained predictive accuracy and generalizability when confronted with new data. For a comprehensive breakdown of the detailed performance metrics, including the validation performance, of the employed models, please refer to Table 10.

Table 10: Detailed Performance Metrics for Selected Models on Validation and Test Sets

Model	Dataset	Recall 0	Recall 1	G-Mean	Macro-Avg F1	AUC
Logistic Regression	Validation	**1.00**	0.00	0.05	0.48	0.656
	Test	**1.00**	0.00	0.05	0.48	0.667
Weighted Logistic Regression	Validation	0.59	0.71	0.64	0.48	0.707
	Test	0.59	0.72	0.65	0.48	0.716
Weighted Decision Tree	Validation	0.87	0.73	0.79	0.69	0.830
	Test	0.86	0.73	0.79	0.69	0.834
Weighted Random Forest	Validation	0.69	**0.83**	0.75	0.57	0.841
	Test	0.69	**0.83**	0.75	0.57	0.839
Weighted Catboost	Validation	0.86	0.82	0.83	0.71	0.923
	Test	0.86	0.81	0.83	0.71	0.922
Weighted XGBoost	Validation	0.88	0.81	**0.84**	**0.72**	**0.925**
	Test	0.88	0.80	**0.84**	**0.72**	**0.924**

5.2.3 *ROC AUC Curves*

The AUC-ROC curves assess the discriminatory performance of different models. As illustrated in the ROC AUC curves (Figure 4), LR exhibits moderate performance with a shallow ascent, indicating limited discriminatory power. WLR improves upon LR, displaying a steeper ascent and higher TPR at similar FPR levels. WRF maintains a strong and stable performance with consistently elevated TPR. However, WDT starts with a higher TPR but faces challenges, as seen in a subsequent drop and a straight-line trajectory, indicating difficulties in balancing TPR and FPR.

The ROC AUC curves for WCB and WXGB demonstrate effective discriminatory performance with steep ascents, emphasizing high TPR at low FPR. Both models show stability and reliability in distinguishing between positive and negative instances.

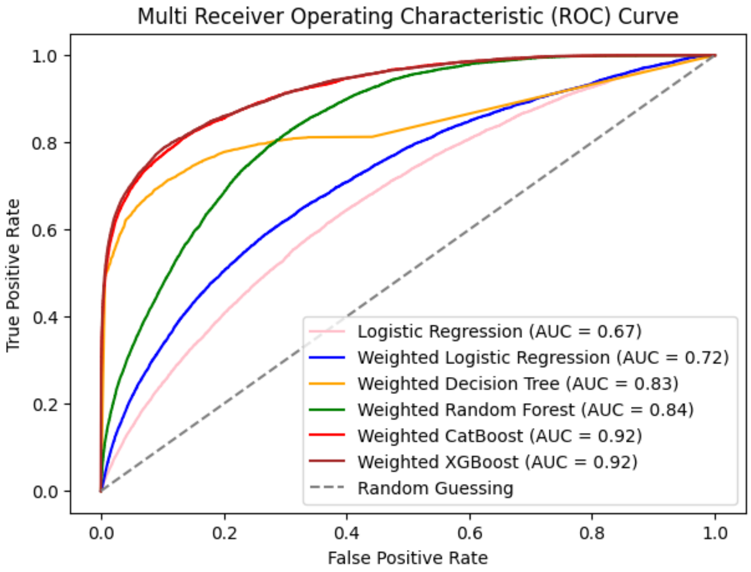


Figure 4: Multi Receiver Operating Characteristic (ROC) curve of selected models

5.2.4 *Log Loss Curve Analysis*

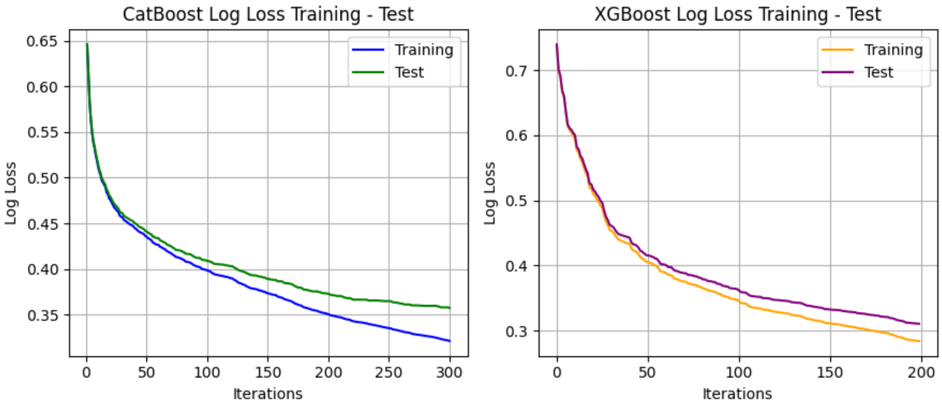


Figure 5: Training and Test Log Loss Curves for Weighted CatBoost and Weighted XGBoost

In Figure 5, the log loss curves of the two best-performing models, WCB and WXGB, are compared side by side. WCB exhibited a gradual decrease in log loss, but towards the end, the training log loss decreased more rapidly than the testing log loss, suggesting potential overfitting. This is indicated by the increasing gap between the two curves. On the other hand, WXGB showed a lower log loss with a more balanced performance, as the testing log loss closely followed the training curve. The smaller gap between the training and testing log loss curves for WXGB suggests effective learning without significant overfitting, highlighting its better generalization capability than WCB.

5.2.5 *Comparison between Models*

In the comparative analysis of ML models for binary classification, distinct characteristics are evident. LR excels in predicting the majority class but exhibits limitations in capturing instances of the minority class, leading to a significant FN rate. Introducing class weights in WLR mitigates this issue, showcasing an improvement in identifying minority class instances. WRF minimizes FNs for the minority class but at the expense of higher misclassifications in the majority class. WDT attains a well-balanced performance, effectively navigating the trade-off between the two classes compared to preceding models. Particularly noteworthy, WCB and WXGB consistently surpass their counterparts, showcasing an enhanced balance in accurately predicting both classes with high precision. These models exhibit proficiency in addressing class imbalance, evidenced by their robust performances across a spectrum of evaluation metrics. This encompasses elevated recall metrics for both classes, a higher geometric mean, macro-average F1 scores, AUC values, and more refined ROC AUC curves compared to alternative models examined in this study.

5.3 *Profitability Analysis*

This subsection is critical in the results, connecting technical evaluations and offering a final, real-world-informed decision on the most effective model.

5.3.1 *Comparative Financial Performance of Models*

The profitability analysis in Table 11 highlights significant performance variations among ML models. LR incurred a notable loss, improved by 737% with class weights in WLR, resulting in a profit of $31,230,786. Ensemble methods, like WRF and WDT, also achieved high-profit increases (837% and 944%, respectively). The most remarkable gains were observed in WCB (989% increase, $43,543,263) and WXGB (1031% increase, $45,595,661). In comparing WCB and WXGB, both models exhibited comparable performance according to model-specific metrics. However, the superior profitability achieved by WXGB in this thesis scenario ultimately designates it as the preferred and most effective model for maximizing financial gains while reducing FNs. For detailed profit values on each model's validation and test sets, please consult Figure 13 in Appendix C (page 45).

Model	**Profit ($)**	**Increase (%)**
Baseline Logistic Regression	-4,897,414	-
Weighted Logistic Regression	31,230,786	737
Weighted Random Forest	36,179,803	837
Weighted Decision Tree	41,372,473	944
Weighted Catboost	43,543,263	989
Weighted XGBoost	45,595,661	1031

Table 11: Profitability Comparison of Selected Models on the Test Set with Percentage Increase (Compared to Baseline Logistic Regression)

5.3.2 *Differential Impact: True Negative Profit and False Negative Loss Across Models*

Figure 6 visually represents model-specific gains from TNs and losses from FNs. LR exhibits extreme TN profits and FN losses, resulting in negative overall profitability. WLR and WRF limit losses from FNs but struggle to maximize TN profits. WDT shows high TN profitability, comparable to WCB and WXGB. WDT's slightly lower performance is due to higher losses from FNs. Comparing WCB and WXGB, their similarly low FN losses highlight their effectiveness. WXGB's superior profitability is attributed to its heightened accuracy in predicting class 0 instances. For a more specific breakdown of each model's average costs associated with FNs and TNs, see Table 15 in Appendix C (page 45).

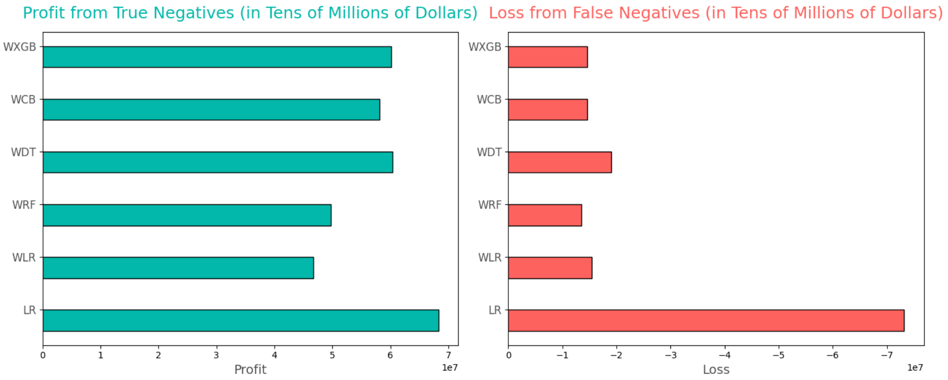
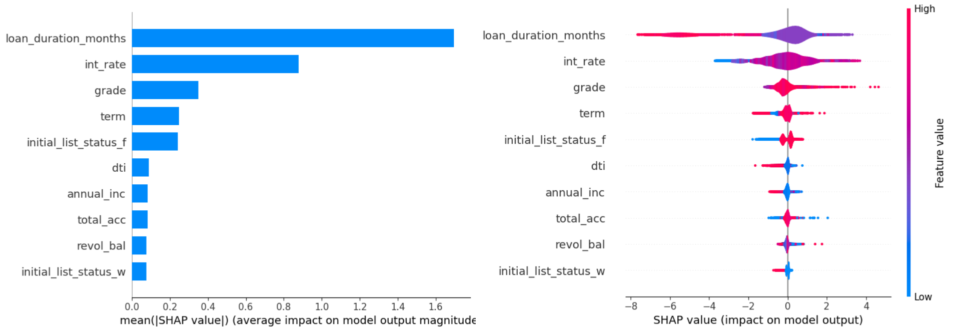


Figure 6: Model-wise Breakdown: Losses from False Negatives and Profits from True Negatives

5.4 *Feature Contribution*

5.4.1 *Global and Individual Feature Contribution*



(a) Global Feature Attribution (b) Individual Feature Contribution

Figure 7: Summary of XGBoost Feature Contributions

In Figure 7a, the x-axis represents the average magnitude change in model output when a feature is integrated. The arrangement of features is determined by the cumulative absolute impact magnitude values they exert on the model (Meng et al., 2020). Key findings highlight `loan_duration_months`, `int_rate`, and `grade` as crucial predictors significantly impacting model outputs. Conversely, features like `term`, `initial_list_status_f`, and `dti` show less pronounced effects, indicating limited influence on predictions. In Figure 7b, the violin SHAP plot unveils the impact of crucial features on the model's binary predictions. Notably, `loan_duration_months` exhibits a distinctive pattern where values towards the left lean the model towards classifying as the negative class, while larger values, especially towards the right, influence the model towards predicting the positive class. This suggests that shorter loan durations contribute more to negative predictions, while longer durations are associated with positive predictions. Similarly, `int_rate` showcases a clear trend where lower interest rates to the far left tend to push the model toward negative predictions. As the interest rate increases, the model leans towards positive predictions. The feature `grade` demonstrates that lower loan grades, positioned towards the left, contribute to negative predictions. In comparison, higher grades strongly influence the model towards positive predictions.

5.4.2 *Feature Ablation*

The detailed feature importance analysis, as depicted in Table 12, offers insights into the performance metrics of the WXGB model when individual features are left out. When `loan_duration_months` is excluded, a substantial and noticeable decline is observed across crucial metrics such as G-mean, AUC, and macro-average F1 scores. This first omission significantly influences the model's ability to distinguish between Class 0 and Class 1 instances. Subsequently, eliminating the `int_rate` feature yields a milder but noticeable reduction in performance metrics and misclassifications in both Class 0 and Class 1. Intriguingly, the third feature, `grade`' displays a further moderated impact, explicitly leading to misclassification within Class 0 instances without drastically affecting the other metrics. The diminishing influence observed in this sequential pattern, moving from the first to the second and subsequently to the third feature, emphasizes their respective order of importance. This observation validates the conclusions presented in subsubsection 5.4.1. Subsequent features exhibit progressively diminishing effects on WXGB performance metrics, highlighting their relative insignificance compared to the initial ones.

Table 12: Weighted XGBoost Performance Metrics with Left-out Features

Left-out Feature	Recall 0	Recall 1	G-mean	Macro-average F1	AUC
No features left out	0.88	0.80	0.84	0.72	0.924
loan_duration_months	0.68	0.64	0.65	0.53	0.719
int_rate	0.74	0.77	0.75	0.59	0.841
grade	0.85	0.80	0.82	0.69	0.915
initial_list_status	0.88	0.80	0.84	0.72	0.924
term	0.88	0.80	0.84	0.72	0.924
installment	0.88	0.80	0.83	0.72	0.924

6 DISCUSSION

6.1 *Analysis of Study Findings and Existing Literature*

While opinions on the best model for predicting loan defaults vary in the literature, a recurring theme highlights the effectiveness of CB and XGB. The results from my study support this trend, affirming that both CB and XGB deliver comparable performances, consistent with existing research. It's worth noting that these findings are in agreement with a specific study (Barbaglia et al., 2023), which further underscores the superior efficacy of XGB.

Moreover, the results underscore the positive impact of incorporating class weighting into the classifier, consistently aligning with insights from other studies (L. Zhu, 2019; M. Zhu et al., 2018), emphasizing the effectiveness of class weighting in achieving elevated performance across diverse classification metrics.

Furthermore, this research affirms three key findings: the efficacy of XGB as a robust model, the positive influence of incorporating class weighting for improved classifier performance, and the synergistic advantages of combining XGB with class weights. These outcomes are in harmony with the broader literature on loan default prediction and are further supported by L. Zhu (2019), endorsing the superiority of XGB when augmented with class weights.

To complement these findings, implementing a custom scoring function led to increased profitability across the models, aligning with a previous study by Serrano-Cinca and Gutierrez-Nieto (2016). These findings underscore the practical efficacy of a profit-driven approach, emphasizing its relevance in decision-making for P2P loan investments. By prioritizing profit over traditional accuracy metrics, this approach not only quantifies the study's outcomes effectively but also highlights its alignment with established research, demonstrating the applicability and utility of such an approach in real-world scenarios where financial gains are paramount.

6.2 *Contributions*

This dissertation significantly advances existing literature by introducing innovative techniques, explicitly employing the underexplored approach of class weighting in cost-sensitive learning and incorporating a tailored scoring function—dimensions that have received limited attention in previous research. The study expands the scholarly debate by investigating state-of-the-art models, namely XGB and CB. It focuses primarily on loan default prediction, specifically emphasizing minimizing FNs. This objective is achieved by implementing class weights and a customized scoring function during the fine-tuning process of these models. The key contributions of this research highlight the crucial role of profit maximization as a targeted outcome, showcasing its profound impact on both the broader economy and individual financial institutions. Notably, the study contends that prioritizing profit maximization over traditional accuracy or precision in model construction can have far-reaching financial implications, providing valuable insights into loan default prediction.

6.3 *Limitations and Challenges*

This study encountered several significant limitations and challenges. For starters, the availability of publicly accessible loan datasets was limited, with barely any available due to ethical and regulatory concerns.

Second, the computational resources required to train the used models posed a significant challenge. CB and XGB were among the models that required significant processing time. Increased computational resources may improve the study's scalability by allowing it to use larger datasets or incorporate additional sources. As a result, more comprehensive hyperparameter tuning and incorporating a broader range of values would be possible.

Moreover, the study deals with the ever-changing nature of the peer-to-peer lending landscape. The models are challenged by the possibility of rapid changes and fluctuations caused by economic conditions. Using historical data for training may only partially capture these dynamic market conditions, emphasizing the importance of continuous adaptation to changing financial landscapes. Given the ever-changing nature of the peer-to-peer lending ecosystem, this dynamic nature poses challenges for the continued relevance and applicability of the employed techniques or models in the present day. On top of that, the incorporation of custom scoring functions designed to maximize profit introduces an additional layer of complexity. Constructing these functions can prove challenging, notably as the costs associated with FNs and TNs may differ and be difficult to represent accurately across various business scenarios. This limitation restricts the practical use of custom scoring functions to individuals with substantial domain knowledge, highlighting the importance of real-world expertise in navigating and implementing practical solutions in this intricate landscape.

6.4 *Future Directions*

6.4.1 *Exploration of Alternative Models*

As the dynamic field of credit risk assessment continues evolving, an opportunity exists to explore alternative models beyond the established frameworks of CB and XGB. Notably, Bhetuwal and Siddanta (2023) proposed a custom neural network in their study, identifying it as the best-performing model in the context of credit risk. Jumaa et al. (2023) proposed using Keras, a TensorFlow neural network library, for loan default prediction, achieving a notably high accuracy. This outcome serves as a conclusive testament to the effectiveness of deep learning in advancing the precision of loan default prediction. Moreover, in another exploration, researchers Bayraci and Susuz (2019) integrated a Deep Neural Network (DNN) based classification model alongside traditional classification methods, specifically in the context of loan default prediction. Their findings emphasized a noteworthy trend where the accuracy of the deep learning classification model exhibited a positive correlation with the size of the dataset. This insight suggests that, especially in scenarios involving extensive and complex datasets related to loan defaults, deep learning models have the potential to outperform ML based models.

6.4.2 *Innovative Approaches to Address Class Imbalance*

Dealing with class imbalance is a critical aspect of credit risk assessment. Future research could delve into innovative techniques for handling imbalanced datasets. Coser et al. (2019), in an extensive study benchmarking oversampling, undersampling, and their combined strategies, revealed that a hybrid approach—specifically, undersampling the majority class and oversampling the minority class could yield superior performance—additionally, techniques such as diversified sensitivity undersampling, as proposed by Y. Chen et al. (2018) could be further employed to enhance the efficacy of handling imbalanced datasets in credit risk assessment.

7 CONCLUSION

Main Research Question:

> *How does integrating class weights to address the class imbalance, along with a profit-maximizing custom scoring function, influence the effectiveness of ML models in minimizing false negatives in the context of loan default prediction?*

Implementing class weights and the custom scoring function reduced raw false negatives across models from the baseline count of 7760. The number diminished notably, falling within the range of 1337-2182 across the models. This considerable improvement in predictive performance for class 1, showcasing enhancements ranging from 71.63% to 82.48% across the models tested, corresponds to a 0.71–0.82 increase in recall for class 1.

Sub-Research Question 1:

> *To what extent does including the custom scoring function and class weights impact the predictive performance of ML models?*

Incorporating class weights and a profit-maximizing custom scoring function resulted in notable changes in model performance compared to the baseline. The models exhibited an increase in G-mean, ranging from 0.6 to nearly 0.8, suggesting an improvement in overall predictive performance. Except for WLR, all models experienced an increase in macro-average F1 by 0.09-0.24. Additionally, there were heightened AUC scores within the range of 0.04-0.25, indicating improved discriminatory ability. Notably, there was an enhanced ability to correctly classify instances for class 1, as indicated by the increased class 1 recall ranging from 0.71 to 0.83.

While these enhancements in overall predictive performance, particularly in correctly predicting class 1 instances, were observed, there was a trade-off. This trade-off manifested in an increased misclassification rate for class 0, with a decrease of 0.12-0.41 compared to the baseline LR. Notably, the top-performing models, WCB and WXGB, demonstrated the most optimal balance by correctly predicting both class 1 and class 0 instances. Compared to the LR baseline, they misclassified a small fraction (0.12-0.14) of class 0 instances while concurrently improving predictive performance for class 0 by a considerable margin (0.80-0.81).

Sub-Research Question 2:

> *Which machine learning model, among those considered, achieves the highest profitability in predicting loan default with class weights and the custom scoring function, emphasizing real-world financial implications?*

The profitability results exhibit a remarkable surge across the models, showing an increase ranging from 737% to 1031% compared to the LR baseline. The WXGB emerges as the top performer, achieving an 80% reduction in false negatives and generating a profit of $45,595,661. In contrast, the baseline model yields a profit of -$4,897,414, representing a 1031% increase in profitability.

REFERENCES

Ali, H., Salleh, M. N. M., Hussain, K., Ahmad, A., Ullah, A., Muhammad, A., Naseem, R., & Khan, M. (2019). A review on data preprocessing methods for class imbalance problem. *International Journal of Engineering & Technology*, *8*, 390–397.

Ali, H., Salleh, M. N. M., Saedudin, R., Hussain, K., & Mushtaq, M. F. (2019). Imbalance class problems in data mining: A review. *Indonesian Journal of Electrical Engineering and Computer Science*, *14*(3), 1560–1571.

Alshari, H., Saleh, A. Y., & Odabas, A. (2021). Comparison of gradient boosting decision tree algorithms for cpu performance. *Journal of Institute of Science and Technology*, *37*(1), 157–168.

Ariza-Garzon, M. J., Arroyo, J., Caparrini, A., & Segovia-Vargas, M.-J. (2020). Explainability of a machine learning granting scoring model in peer-to-peer lending. *Ieee Access*, *8*, 64873–64890.

Aslam, U., Tariq Aziz, H. I., Sohail, A., & Batcha, N. K. (2019). An empirical study on loan default prediction models. *Journal of Computational and Theoretical Nanoscience*, *16*(8), 3483–3488.

Bachmann, A., Becker, A., Buerckner, D., Hilker, M., Kock, F., Lehmann, M., Tiburtius, P., & Funk, B. (2011). Online peer-to-peer lending-a literature review. *Journal of Internet Banking and Commerce*, *16*(2), 1.

Barbaglia, L., Manzan, S., & Tosetti, E. (2023). Forecasting loan default in europe with machine learning. *Journal of Financial Econometrics*, *21*(2), 569–596.

Basel Committee on Banking Supervision. (2000). *Principles for the management of credit risk – final document* (Final Document). Bank for International Settlements (BIS).

Bayraci, S., & Susuz, O. (2019). A deep neural network (dnn) based classification model in application to loan default prediction. *Theoretical and Applied Economics*, *26*(4(621)), 75–84.

Bengio, Y., & Grandvalet, Y. (2003). No unbiased estimator of the variance of k-fold cross-validation. *Advances in Neural Information Processing Systems*, *16*.

Bhetuwal, A., & Siddanta, K. (2023). *Loan default prediction and comparison of various machine learning models* [MS Thesis]. Katz School of Science and Health, Yeshiva University.

Cao, J., Lu, H., Wang, W., & Wang, J. (2013). A loan default discrimination model using cost-sensitive support vector machine improved by pso. *Information Technology and Management*, *14*, 193–204.

Chacon, S., & Straub, B. (2014). *Pro git*. Apress.

Chen, N., Ribeiro, B., & Chen, A. (2016). Financial credit risk assessment: A recent review. *Artificial Intelligence Review*, *45*, 1–23.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chen, Y., Leu, J., Huang, S., Wang, J., & Takada, J. (2021). Predicting default risk on peer-to-peer lending imbalanced datasets. *IEEE Access*, *9*, 73103–73109.

Chen, Y., Zhang, J., & Ng, W. W. (2018). Loan default prediction using diversified sensitivity undersampling. *2018 International Conference on Machine Learning and Cybernetics (ICMLC)*, *1*, 240–245.

Coser, A., Maer-Matei, M. M., & Albu, C. (2019). Predictive models for loan default risk assessment. *Economic Computation & Economic Cybernetics Studies & Research*, *53*(2).

Disha, R. A., & Waheed, S. (2022). Performance analysis of machine learning models for intrusion detection system using gini impurity-based weighted random forest (giwrf) feature selection technique. *Cybersecurity*, *5*(1), 1.

Dorogush, A. V., Ershov, V., & Gulin, A. (2017). Catboost: Gradient boosting with categorical features. *Workshop on ML Systems at NIPS 2017*.

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

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

Fauzan, M. A., & Murfi, H. (2018). The accuracy of xgboost for insurance claim prediction. *Int. J. Adv. Soft Comput. Appl*, *10*(2), 159–171.

Federal Reserve Bank of New York Research and Statistics Group. (2023). *Analysis based on new york fed consumer credit panel/equifax data* (tech. rep.). Federal Reserve Bank of New York.

Feng, W., Huang, W., & Ren, J. (2018). Class imbalance ensemble learning based on the margin theory. *Applied Sciences*, *8*(5), 815.

Fernandez, A., Garcia, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Algorithm-level approaches. *Learning from Imbalanced Data Sets*, 123–146.

Fernandez, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, *61*, 863–905.

Fushiki, T. (2011). Estimation of prediction error by using k-fold cross-validation. *Statistics and Computing*, *21*, 137–146.

Gramegna, A., & Giudici, P. (2021). Shap and lime: An evaluation of discriminative power in credit risk. *Frontiers in Artificial Intelligence*, *4*, 752558.

Gray, K. S. (1985). Can student loan default be forecast accurately? *Journal of Student Financial Aid*, *15*(1), Article 3.

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, *35*, 507–520.

Guo, X., Yin, Y., Dong, C., Yang, G., & Zhou, G. (2008). On the class imbalance problem. *2008 Fourth international conference on natural computation*, *4*, 192–201.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Rio, J. F., Wiebe, M., Peterson, P., . . . Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362.

Harvey, C. R., & Liu, Y. (2020). False (and missed) discoveries in financial economics. *The Journal of Finance*, *75*(5), 2503–2553.

Hido, S., Kashima, H., & Takahashi, Y. (2009). Roughly balanced bagging for imbalanced data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *2*(5-6), 412–426.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95.

Ibrahim, A. A., Ridwan, R. L., Muhammed, M. M., Abdulaziz, R. O., & Saheed, G. A. (2020). Comparison of the catboost classifier with other machine learning methods. *International Journal of Advanced Computer Science and Applications*, *11*(11).

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1), 1–54.

Jorgensen, T. (2018). Peer-to-peer lending-a new digital intermediary, new legal challenges. *NJCL*, 231.

Jumaa, M., Saqib, M., & Attar, A. (2023). Improving credit risk assessment through deep learning-based consumer loan default prediction model. *International Journal of Finance & Banking Studies*, *12*(1), 85–92.

Krichene, A. (2017). Using a naive bayesian classifier methodology for loan risk assessment: Evidence from a tunisian commercial bank. *Journal of economics, finance and administrative science*, *22*(42), 3–24.

Langley, P., & Simon, H. A. (1995). Applications of machine learning and rule induction. *Commun. ACM*, *38*(11), 54–64.

Leevy, J. L., Hancock, J., Zuech, R., & Khoshgoftaar, T. M. (2021). Detecting cybersecurity attacks across different network features and learners. *Journal of Big Data*, *8*(1), 1–29.

Lemaitre, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, *18*(17), 1–5.

Lenz, R. (2016). Peer-to-peer lending: Opportunities and risks. *European Journal of Risk Regulation*, *7*(4), 688–700.

Liu, X.-Y., & Zhou, Z.-H. (2006). The influence of class imbalance on cost-sensitive learning: An empirical study. *sixth international conference on data mining (ICDM'06)*, 970–974.

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*.

Madaan, M., Kumar, A., Keshri, C., Jain, R., & Nagrath, P. (2021). Loan default prediction using decision trees and random forest: A comparative study. *IOP Conference Series: Materials Science and Engineering*, *1022*(1), 012042.

Meng, Y., Yang, N., Qian, Z., & Zhang, G. (2020). What makes an online review more helpful: An interpretation framework using xgboost and shap values. *Journal of Theoretical and Applied Electronic Commerce Research*, *16*(3), 466–490.

Moula, F. E., Guotai, C., & Abedin, M. Z. (2017). Credit default prediction modeling: An application of support vector machine. *Risk Management*, *19*, 158–187.

Mukid, M., Widiharih, T., Rusgiyono, A., & Prahutama, A. (2018). Credit scoring analysis using weighted k nearest neighbor. *Journal of Physics: Conference Series*, *1025*(1), 012114.

Naganjaneyulu, S., & Kuppa, M. R. (2013). A novel framework for class imbalance learning using intelligent under-sampling. *Progress in Artificial Intelligence*, *2*, 73–84.

Namvar, A., Siami, M., Rabhi, F., & Naderpour, M. (2018). Credit risk prediction in an imbalanced social lending environment. *arXiv preprint arXiv:1805.00801*.

pandas development team, T. (2020, February). *Pandas-dev/pandas: Pandas* (Version 1.5.3). Zenodo.

Park, E., Ahn, J., & Yoo, S. (2017). Weighted-entropy-based quantization for deep neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5456–5464.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Phan, T. H., & Yamamoto, K. (2020). Resolving class imbalance in object detection with weighted cross entropy losses. *arXiv preprint arXiv:2006.01413*.

Ponsam, J. G., Gracia, S. J. B., Geetha, G., Karpaselvi, S., & Nimala, K. (2021). Credit risk analysis using lightgbm and a comparative study of popular algorithms. *2021 4th International Conference on Computing and Communications Technologies (ICCCT)*, 634–641.

Priscilla, C. V., & Prabha, D. P. (2020). Influence of optimizing xgboost to handle class imbalance in credit card fraud detection. *2020 third international conference on smart systems and inventive technology (ICSSIT)*, 1309–1315.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). Catboost: Unbiased boosting with categorical features. *Advances in neural information processing systems*, *31*.

Prusty, S., Patnaik, S., & Dash, S. K. (2022). Skcv: Stratified k-fold cross-validation on ml classifiers for predicting cervical cancer. *Frontiers in Nanotechnology*, *4*, 972421.

Rahman, M. M., & Davis, D. N. (2013). Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing*, *3*(2), 224.

Rezvani, S., & Wang, X. (2023). A broad review on class imbalance learning techniques. *Applied Soft Computing*, 110415.

R.G. (2021). Credit risk analysis.

Riyanto, S., Imas, S. S., Djatna, T., & Atikah, T. D. (2023). Comparative analysis using various performance metrics in imbalanced data for multi-class text classification. *International Journal of Advanced Computer Science and Applications*, *14*(6).

Saber, M., Boulmaiz, T., Guermoui, M., Abdrabo, K. I., Kantoush, S. A., Sumi, T., Boutaghane, H., Nohara, D., & Mabrouk, E. (2022). Examining lightgbm and catboost models for wadi flash flood susceptibility prediction. *Geocarto International*, *37*(25), 7462–7487.

Schlomer, N., danielhkl, Wehrfritz, A., Berndt, H., Stathopoulos, S., Boeddeker, C., Edler, D., Spott, A., Gaul, A., Rossi, M., Vinot, B., Schürmann, D., Lipp, M., Dawson, D., mrtnschltr, pwohlhart, hgwd2, Koslowski, S., Lacasse, P., . . . Kuzmin, A. (2018, February). *Nschloe/matplotlib2tikz v0.6.15* (Version v0.6.15). Zenodo.

Serrano-Cinca, C., & Gutierrez-Nieto, B. (2016). The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (p2p) lending. *Decision Support Systems*, *89*, 113–122.

Shingi, G. (2020). A federated learning based approach for loan defaults prediction. *2020 International Conference on Data Mining Workshops (ICDMW)*, 362–368.

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

Van Rossum, G., & Drake Jr, F. L. (1995). *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021.

Weiss, G. M., & Provost, F. (2003). Learning when training data are costly: The effect of class distribution on tree induction. *Journal of artificial intelligence research*, *19*, 315–354.

Zhu, L. (2019). *Predictive modelling for loan defaults*. University of California, Los Angeles.

Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. *Procedia Computer Science*, *162*, 503–513.

Zhu, M., Xia, J., Jin, X., Yan, M., Cai, G., Yan, J., & Ning, G. (2018). Class weights random forest algorithm for processing class imbalanced medical data. *IEEE Access*, *6*, 4641–4652.

APPENDIX A

Table 13: Dataset Feature Description

Feature	**Description**
loan_amnt	Loan amount
term	Number of payments
int_rate	Interest rate
installment	Monthly payment
grade	LC assigned loan grade
emp_length	Employment length
home_ownership	Home ownership status
annual_inc	Annual income
verification_status	Verification status
loan_status	Loan status
purpose	Category for the loan request
dti	Monthly debt payments on the total debt obligations
open_acc	Number of open credit lines
pub_rec	Number of derogatory public records
revol_bal	Total credit revolving balance
revol_util	Revolving line utilization rate
total_acc	Total number of credit lines
initial_list_status	Initial listing status
loan_duration_months	Loan duration
application_type	Indicates whether the loan is an individual or a joint application

Table 14: Missing Values in the Dataset

Feature	Missing Values
loan_amnt	0
term	0
int_rate	0
installment	0
grade	0
emp_length	44,825
home_ownership	0
annual_inc	4
verification_status	0
loan_status	0
purpose	0
dti	0
open_acc	29
pub_rec	29
revol_bal	0
revol_util	502
total_acc	29
initial_list_status	0
loan_duration_months	17,659
application_type	0

APPENDIX B

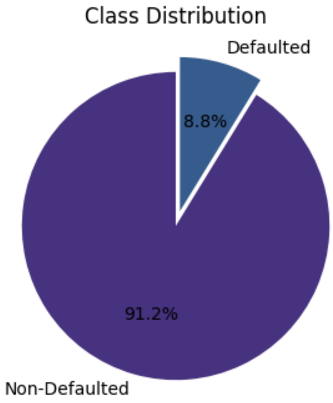


Figure 8: Target Class Distribution: Defaulted vs. Non-Defaulted

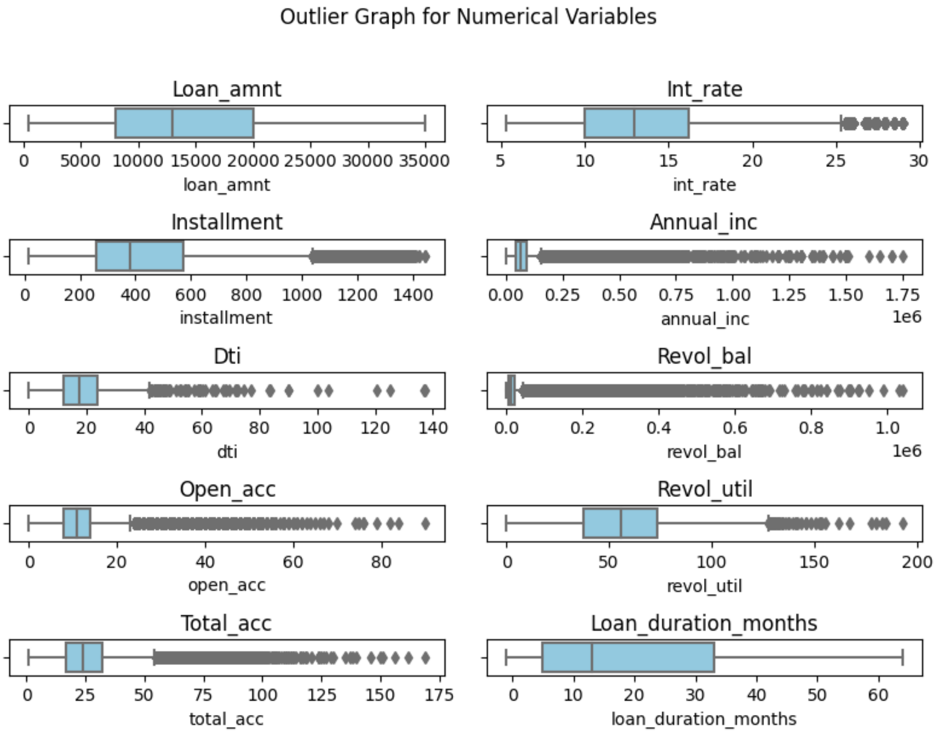
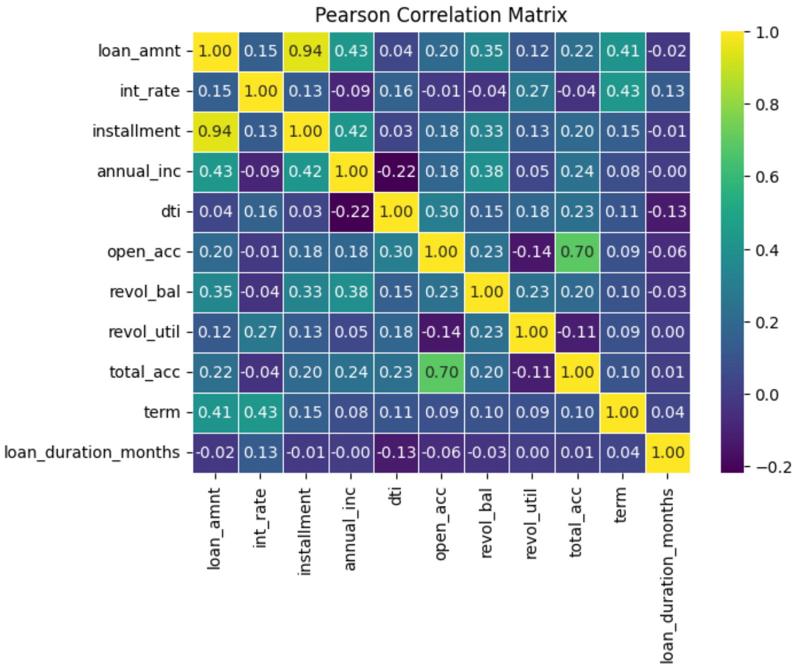


Figure 9: Outlier Boxplots of Numerical Predictors



	loan_amnt	int_rate	installment	annual_inc	dti	open_acc	revol_bal	revol_util	total_acc	term	loan_duration_months
loan_amnt	1.00	0.15	0.94	0.43	0.04	0.20	0.35	0.12	0.22	0.41	-0.02
int_rate	0.15	1.00	0.13	-0.09	0.16	-0.01	-0.04	0.27	-0.04	0.43	0.13
installment	0.94	0.13	1.00	0.42	0.03	0.18	0.33	0.13	0.20	0.15	-0.01
annual_inc	0.43	-0.09	0.42	1.00	-0.22	0.18	0.38	0.05	0.24	0.08	-0.00
dti	0.04	0.16	0.03	-0.22	1.00	0.30	0.15	0.18	0.23	0.11	-0.13
open_acc	0.20	-0.01	0.18	0.18	0.30	1.00	0.23	-0.14	0.70	0.09	-0.06
revol_bal	0.35	-0.04	0.33	0.38	0.15	0.23	1.00	0.23	0.20	0.10	-0.03
revol_util	0.12	0.27	0.13	0.05	0.18	-0.14	0.23	1.00	-0.11	0.09	0.00
total_acc	0.22	-0.04	0.20	0.24	0.23	0.70	0.20	-0.11	1.00	0.10	0.01
term	0.41	0.43	0.15	0.08	0.11	0.09	0.10	0.09	0.10	1.00	0.04
loan_duration_months	-0.02	0.13	-0.01	-0.00	-0.13	-0.06	-0.03	0.00	0.01	0.04	1.00

Figure 10: Numerical Pearson Correlation Matrix

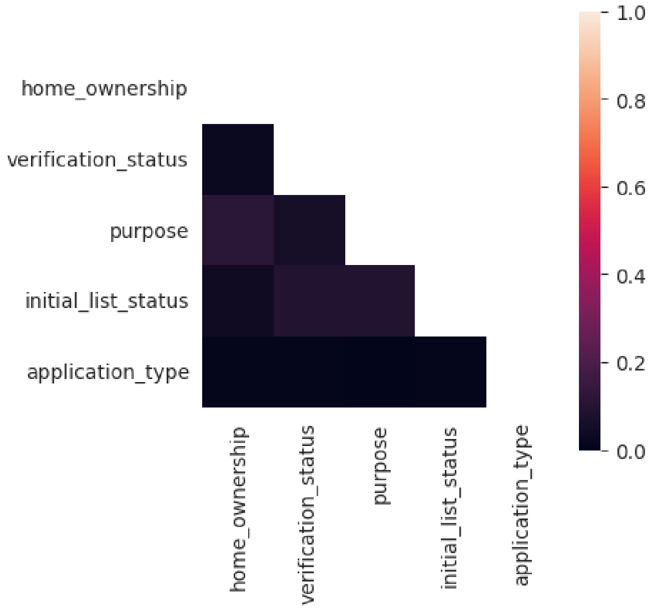


Figure 11: Cramér's V Correlation Matrix of Nominal Predictors

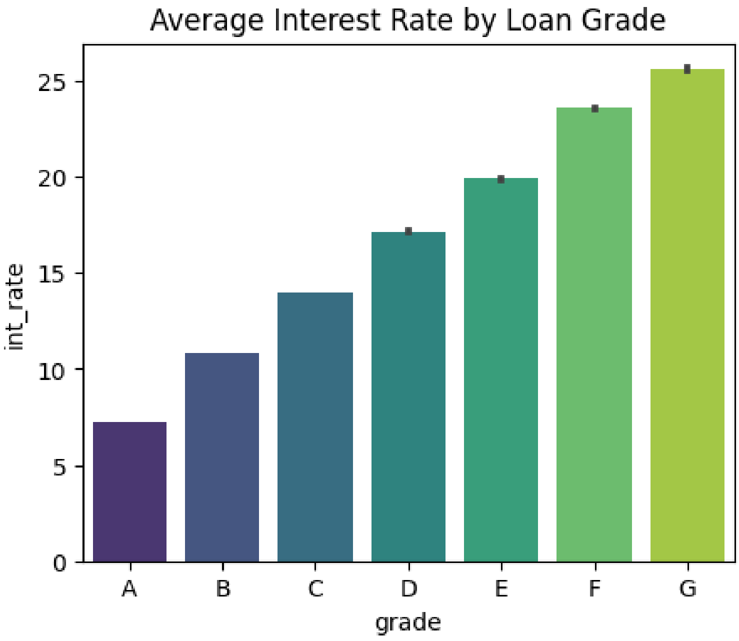


Figure 12: Average Interest Rate by Loan Grade

APPENDIX C

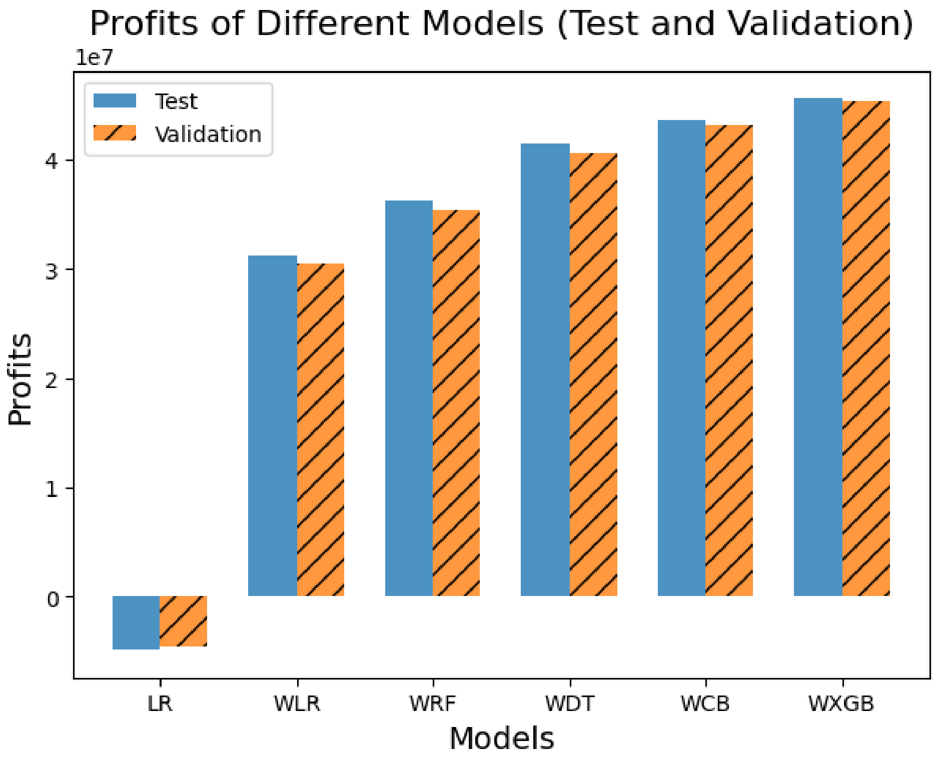


Figure 13: Comparison of Model-Wise Profits on Validation and Test Datasets. Legends: LR (Logistic Regression), WLR (Weighted Logistic Regression), WDT (Weighted Decision Tree), WRF (Weighted Random Forest), WCB (Weighted CatBoost), WXGB (Weighted XGBoost)

Table 15: Average Costs of True and False Negatives for Selected Models

Model	**Avg Cost of True Negatives ($)**	**Avg Cost of False Negatives ($)**
Logistic Regression	844.79	-9,437.868
Weighted Logistic Regression	981.428	-7,088.692
Weighted Decision Tree	863.023	-9,154.520
Weighted Random Forest	888.416	-10,112.173
Weighted CatBoost	831.504	-9,924.093
Weighted XGBoost	847.382	-9,589.921

Note: The average cost of True Negatives (TNs) and False Negatives (FNs) is in dollars and was calculated by taking the total profit from TNs and the total loss from FNs, and dividing each by their respective number.