MASTER

MATHEMATICAL FINANCE

MASTER'S FINAL WORK

INTERNSHIP REPORT

ADVANCED MACHINE LEARNING METHODS IN DEVELOPING CREDIT SYSTEMS

LUÍS CARVALHO MORAIS ROSA

SUPERVISION:

PROFESSOR JOÃO AFONSO BASTOS, PhD

DR. PEDRO LOUREIRO

JUNE – 2025

GLOSSARY

CatBoost Categorical Boosting. i, ii, 1–4, 9, 15–19, 22, 25–27

GBDT Gradient Boosting Decision Tree. i

LightGBM Light Gradient Boosting Machine. i, ii, 1–4, 9, 16–19, 22, 25–27

ML Machine Learning. i, 1, 27

XGBoost Extreme Gradient Boosting. i, ii, vi, 1–4, 9, 12, 13, 17, 18, 25, 26

Abstract and Keywords

This thesis comprehensively analyzes the development and evaluation of credit risk prediction models within the Angolan financial system, characterized by substandard data quality and significant challenges in data acquisition and consistency maintenance. The analysis examines the pressing necessity for robust, reliable, and comprehensible credit risk assessment methodologies that facilitate financial inclusion and effective risk management in developing market environments.

The empirical investigation employs an extensive dataset of actual mortgages from the Angolan banking sector. The dataset has numerous issues, including absent values, incomplete credit histories, and significant class disparity in default rates. These issues illustrate the deficiencies in Angola's financial system regarding its infrastructure and reporting mechanisms. A stringent data preprocessing pipeline was established to address these issues. The process encompassed median imputation for absent data, the creation of indicator flags for problematic characteristics, and improved management of infinite and erratic values. To address class imbalance and enhance the model's generalization to infrequent default events, we investigated resampling techniques such as SMOTE, SMOTE-Tomek, and ADASYN.

The modeling system encompassed both conventional Logistic regression and sophisticated tree-based ensemble methods, including Categorical Boosting (CatBoost), Light Gradient Boosting Machine (LightGBM), and Extreme Gradient Boosting (XGBoost). We examined each model both with and without resampling techniques, enabling a rigorous analysis of the interplay between model complexity, data augmentation, and predictive performance in the context of low-quality data. We focused on calibrating the model by isotonic regression to enhance the accuracy of risk assessments, essential for informed regulatory financial decisions.

The findings indicate that CatBoost and LightGBM exhibit superior discrimination capabilities (assessed using ROC-AUC and PR-AUC) compared to linear models, even in the presence of imperfect data. Nonetheless, Logistic regression retained its utility in contexts where comprehension and auditability were paramount. The inquiry indicates that contemporary gradient boosting algorithms may autonomously address data imbalance, rendering synthetic resampling less advantageous in high-capacity models.

This thesis demonstrates the significance of data quality for the efficacy and reliability of credit scoring systems in Angola. It demonstrates that machine learning frameworks, when integrated with sophisticated preprocessing and calibration techniques, can significantly enhance forecast accuracy and operational reliability. The findings endorse the continuous enhancement of Angola's data infrastructure and indicate that ensemble-based model stacking may enhance the reliability and consistency of future risk assessment initiatives.

KEYWORDS: Credit risk assessment; Mortgage default prediction; Angolan financial system; Data quality; Imbalanced data; Machine learning; CatBoost; LightGBM; Logistic Regression;

Table of Contents

LIST OF FIGURES

List of Tables

ACKNOWLEDGEMENTS

First and foremost, I would like to pay tribute to my family, especially my parents - Luís and Aurora - for the nurturing environment they created for my development and education. Curiosity and the pursuit of knowledge have always been the key drivers of my learning, harmoniously combined with the freedom to think, express, and develop any idea. To them, I dedicate this document and express my eternal gratitude for their unwavering support, investment, and prioritization of my education.

I am also deeply grateful to all the professors and educators who have guided me throughout my academic journey. In particular, I wish to thank Professor João Afonso Bastos for his invaluable guidance and advice during the preparation of this thesis, and Professor João Janela for providing a safe and serene space for orientation and support.

At KPMG, I extend my special thanks to Pedro Loureiro for the opportunity and kindness he has shown me, and to Ricardo Soares for generously sharing his knowledge and offering continuous guidance.

I would like to acknowledge my fellow Master's colleagues— Mariana Silva, Elisa Alves, Catarina Neves, and Guilherme Gonçalves. Sharing this journey with you made every step twice as enjoyable and memorable.

Lastly, my sincere thanks to Simão Alves, Gabriel Fonseca and Tomás Bauwens for their constant support and friendship. Your presence provided a much-needed space to reflect, recharge, and find balance throughout this academic journey.

Thank you all for being part of this important chapter of my life.

1 Introduction

This thesis is a direct outcome of my internship in the Risk department of KPMG Portugal, where I focused on credit risk modeling and examined the challenges faced by financial institutions in developing robust credit risk assessment systems, particularly in emerging economies such as Angola. During this period, I encountered practical challenges such as data asymmetry, low data quality, absent credit histories, and the constraints of conventional modeling techniques. These challenges provided me with invaluable practical experience and ignited my desire to explore innovative solutions.

During my tenure at KPMG, I observed how sophisticated Machine Learning (ML) methodologies may transform credit risk assessment and facilitate access to financial services. This experience culminated in this thesis, which is a comprehensive analysis and comparison of contemporary modeling techniques specifically tailored to the distinctive characteristics of Angolan credit data.

Access to credit is often seen as a crucial determinant of economic growth, since it enables individuals and enterprises to invest, expand, and manage financial risks. In developing nations such as Angola, the implementation of effective credit scoring systems is often hindered by inadequate financial infrastructure, incomplete credit histories, and persistent data quality issues. These factors exacerbate financial exclusion, particularly for underserved communities, and hinder bank's ability to accurately assess and manage credit risk.

Logistic regression and other conventional credit scoring models have long been the industry standard due to their comprehensibility and regulatory acceptance. Nevertheless, these models frequently struggle to accurately represent the intricate, non-linear relationships present in actual financial data, particularly when the datasets are extensive and diverse. Recent advancements in ML have resulted in robust alternatives such as CatBoost, LightGBM, and XGBoost, which exhibit superior predictive capabilities and enhanced flexibility. These ensemble methods excel in handling complex data structures and have demonstrated potential in enhancing credit risk assessment across several industries.

A significant issue with employing ML models for credit scoring in Angola is the pronounced imbalance in credit datasets, characterized by a minimal amount of default instances. This discrepancy may cause models to favor the majority class, so complicating the identification of high-risk applicants. Absence of data, outliers, and erroneous records are among the challenges that complicate modeling significantly. These factors may diminish the reliability and accuracy of models. To address these issues, it is essential to deal with low quality records, as they may compromise the model's reliability and precision. To address these issues, a comprehensive data preparation strategy is required, incorporating advanced imputation techniques, outlier capping, in addition to that, test of resampling methods use such as SMOTE, SMOTETomek, and ADASYN will be done.

The modeling process requires feature selection and model calibration. To enhance model comprehensibility, reduce dimensionality, and mitigate multicollinearity, we employ techniques such as correlation analysis, Variance Inflation Factor (VIF) assessment, and SHAP value interpretation. Calibration methods, including isotonic regression, ensure that predicted probabilities accurately reflect actual default risk. This is essential for effective risk management and compliance with regulations.

This research utilizes an extensive dataset of Angolan credit information to systematically evaluate the efficacy of Logistic Regression, Stepwise selection, CatBoost, LightGBM and XGBoost in predicting credit risk. The research employs advanced feature selection, rigorous cross-validation, and probability calibration at each stage of the modeling process. Evaluation metrics extend beyond mere accuracy to encompass precision, recall, F1-score, ROC-AUC, PR-AUC, and Brier Score, providing a comprehensive assessment of a model's performance.

This thesis aims to provide valuable recommendations for enhancing credit risk assessment in developing nations by examining the particular issues of data imbalance and quality in Angola. The ultimate objective is to enhance access to financial services for a greater number of individuals and to foster sustained economic growth.

2 LITERATURE REVIEW

Credit scoring plays a crucial role in the credit system development, which works as a reliable support in financial decision-making, as an example of a developing credit system, the African credit market faces distinct institutional and structural challenges that impede traditional credit scoring approaches. Andrianova et al.(2010) propose a "bad credit market equilibrium" framework, attributing financial underdevelopment to moral hazard and adverse selection. Their empirical analysis of African banks reveals a paradox: despite excess liquidity, lending remains limited, suggesting that structural issues, not savings mobilization, constrain credit markets. Poor regulatory quality exacerbates loan defaults, although threshold effects indicate diminishing returns beyond a certain regulatory level. These findings stress the need for credit scoring models that can adapt to such complexities, as conventional methods may not adequately address these structural findings.

Part of traditional credit risk approach in this context, Logistic regression, working as a baseline method, assumes linearity between independent variables and the log-odds of default, limiting its effectiveness with non-linear relationships. Song(2025) highlights that financial data, with variables like income and payment history interacting non-linearly, often exceeds logistic regression's capacity. This limitation is acute in developing markets, where diverse borrower profiles and economic volatility increase data complexity. Moreover, Logistic regression struggles with imbalanced datasets, common in credit scoring where defaults are rare, further reducing its predictive power Chen(2024).

As a possible solution, ensemble models, particularly gradient boosted decision trees as LightGBM and XGBoost, may address the limitations of traditional methods by capturing complex, non-linear relationships in credit data. Bastos(2018) demonstrates that these models, which aggregate multiple decision trees via weighted voting, outperform multilayer perceptrons and support vector machines across various metrics. Their ability to classify attributes driving default risk improves interpretability, helping decision making. Regularization ensures robustness against overfitting, making ensemble models ideal for real-time credit scoring in complex markets Bastos & Matos(2021).

CatBoost, developed by Dorogush et al.(2018), excels in credit scoring due to its efficient handling of categorical variables, prevalent in financial datasets. Unlike other boosting algorithms requiring extensive preprocessing, CatBoost's encoding method minimizes overfitting. Xia et al.(2024) integrate the model with large language models and FocalPoly loss to tackle high credit risk and class imbalance in fintech lending, achieving superior performance. Its ordered boosting technique further enhances accuracy by mitigating target leakage, making CatBoost particularly suitable for diverse borrower profiles where categorical data is common.

LightGBM, introduced by Ke et al.(2017), enhances gradient boosting efficiency through Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). GOSS prioritizes instances with larger gradients, while EFB reduces feature dimensionality, enabling LightGBM to train up to 20 times faster than XGBoost with comparable accuracy. Its histogram-based approach and leaf-wise tree growth excel with sparse, high-dimensional credit datasets, common in financial applications. These attributes make this model a scalable solution for credit scoring, where large datasets demand computational efficiency.

XGBoost, a robust gradient boosting algorithm, leverages regularization and missing data handling to improve credit scoring. Zedda(2024) compares XGBoost to logistic regression across a large dataset of Italian SME's, finding sector-specific performance advantages. While XGBoost excels in capturing non-linear interactions, its practical benefits depend on cutoff settings and calibration. This variability suggests that the implementation must be tailored to local market dynamics, balancing predictive power with operational applicability.

Accurate model calibration ensures that predicted probabilities align with actual default rates, a critical requirement in credit scoring. Gneiting & Raftery(2007) established strictly proper scoring rules, with the Brier Score, measuring mean squared differences between predictions and outcomes, emerging as a key metric for assessing both discrimination and calibration. Studies show that isotonic regression improves long-term calibration in time series credit data, supporting risk-adjusted pricing and regulatory compliance Fonseca & Lopes(2017). In developing markets, where regulatory oversight is evolving, calibration is crucial to avoid capital penalties under internal ratings-based approaches and miscalculation of the underlying default probabilities.

3 METHODOLOGY

3.1 Data Description

The dataset utilized in this study originates from a private bank in Angola and covers the period from January 2021 to December 2024. The data is categorized by client and month, which includes a total of 3,895 private clients. Displaying 106,716 monthly positions in a specific bank across main credit types: Housing credit, car loans, and personal credit. The data set includes many critical parameters to assess credit risk, customer behavior, and demographic characteristics.

TABLE I: VARIABLE LIST

Variable Name	**Description**
Year	Year of observation
Month	Month of observation
Artificial Id	Unique identifier for each client
Exposure	Removed due to data leakage
Impairment	Removed due to data leakage
Net Allocated Value	Net allocated value to the collateral
Colateral	Type of collateral
Default Flag	Binary indicator of default
Average Balance	Average account balance
Overdue Credit	Amount of overdue credit
Account Other Bank	Indicator for accounts held also at other banks
Credit Count	Number of credit products held
Solely Guarantors	Indicator for clients only as guarantors
Housing Cred	Indicator for housing credit
Auto Personal Margin Cred	Auto, personal, or margin credit indicator
Education Level	Education level (Angolan classification)
Collaborator or Protocol	Collaboration/protocol with affiliated companies
Employment Sector	Demographic labour indicator
Clients Assets	Value of client assets

Multiple categorical variables were encoded to facilitate the modeling process and enhance the comprehensibility of the results. Employing a hybrid approach that integrated Weight of Evidence (WoE) analysis with expert judgment to encode and prioritize the variables Type of Collateral and Employment Sector. This provided us with risk-based rankings for both categories, while the Net Allocated Value indicated the value and status of the collateral.

An ordinal technique was used to encode educational levels, indicating higher degrees of education. These encoding approaches ensure that categorical variables are represented in a manner that is both statistically valid and aligned with domain expertise.

Certain variables were omitted to prevent data leakage and enhance model performance. Removing the variables Exposure and Impairment from the dataset as they were closely associated with the Dependent variable, thereby distorting the modeling process. Furthermore, to enhance the visibility of operational statuses within the dataset, multiple indicators were transformed into binary flags, such as Collaborator or Protocol.

The primary outcome variable in this study is the Default Flag, a binary indicator of credit default. A number of 1 indicates default, while a value of 0 indicates non-default. The distribution of this characteristic is markedly unequal. There are 103,578 non-default observations and 3,138 default observations, constituting 97.1% and 2.9% of the dataset, respectively. The dataset possesses numerous variables useful to predict credit default.

We developed metrics for temporal insights and variation rates, encompassing av_bal_3m, av_bal_6m, av_bal_9m, and av_bal_12m. This illustrates the average client balances across 3, 6, 9, and 12 months and their temporal variations. Variables such as Overdue Credit and Credit Count indicate client utilization and repayment of credit. Client attributes such as Client Assets, Encoded Employment Sector, and Encoded Educational Level provide further demographic and socio-economic insights.

A significant issue with the dataset is the absence of certain values, particularly in variables requiring historical data over an extended time frame. The variable Client Assets contains 39,274 missing entries, while the temporal balance metrics av_bal_12m, av_bal_9m, av_bal_6m, and av_bal_3m exhibit missing values of 25,177, 16,481, 11,840, and 6,414 respectively. This pattern indicates that the volume of absent data increases with an extended observation time. This is common in financial datasets where recent clients possess abbreviated transaction histories.

In addition to absent data, other variables exhibit significant positive skewness. The skewness values for Average Balance, Net Allocated Value, and Client Assets vary from 13.72 to 26.21. This degree of skewness may alter the results of parametric models and complicate the assessment of the significance of specific traits. The variation rate metrics, av_bal_var_rate_3m, av_bal_var_rate_6m, av_bal_var_rate_9m, and av_bal_var_rate_12m, exhibit several infinite values (ranging from 147 to 215 per variable) due to instances of division by zero.

All these issues will be addressed in Subsection 2.2.

3.2 Data Preprocessing

The preprocessing of the data for this study addressed specific issues, including several missing values, significant positive skewness, and class imbalance. Implementing the subsequent measures to ensure the model functioned effectively and remained comprehensible, while preserving the economic significance of the underlying variables. To prevent data leakage, all preparatory procedures were conducted only on the training set.

Due to the significant skewness of the financial variables in this dataset, median imputation was used as the primary method for addressing missing values. Median imputation is particularly effective in the presence of outliers and skewed distributions, as it is less influenced by extreme values compared to mean imputation. This method preserves the central tendency of the data and the integrity of the distribution, which are crucial for further modeling. Since non-standard data structures are frequent in financial statistics.

Creating box plots of the principal continuous variables prior to and following winsorization allows to examine their distribution, skewness, and the presence of outliers. Prior to winsorization, Figure 1 illustrates the distribution of the primary continuous variables. The image illustrates a distinctly right-skewed distribution in variables such as Average_Balance, Net_Allocated_Value, and particularly Client_Assets. This is evidenced by the extended upper whiskers and the presence of numerous outlier spots much above the median. This phenomenon is prevalent in financial statistics, where a few number of clients own balances or assets significantly beyond the average, resulting in pronounced positive skewness.

Figure 2 illustrates a significant reduction in data dispersion and the influence of extreme outliers following the application of asymmetric winsorization at the 5th and 90th percentiles.

The distribution narrows, resulting in fewer outliers, and the whiskers significantly shorten. This modification preserves the majority of the data's structure while diminishing the impact of outliers, which is crucial for precise model estimation and interpretation. The pronounced positive skewness observed in the data suggests the application of asymmetric caps. This enables the management of the the upper tail more assertively while maintaining the lower range of the data.

Subsequent to Winsorization, the RobustScaler was employed to normalize all continuous variables. The RobustScaler use the median and interquartile range (IQR) for centering and scaling, rendering it very resistant to outliers and suitable for skewed data. This method preserves the data structure while ensuring that the scaled features are uniformly sized.

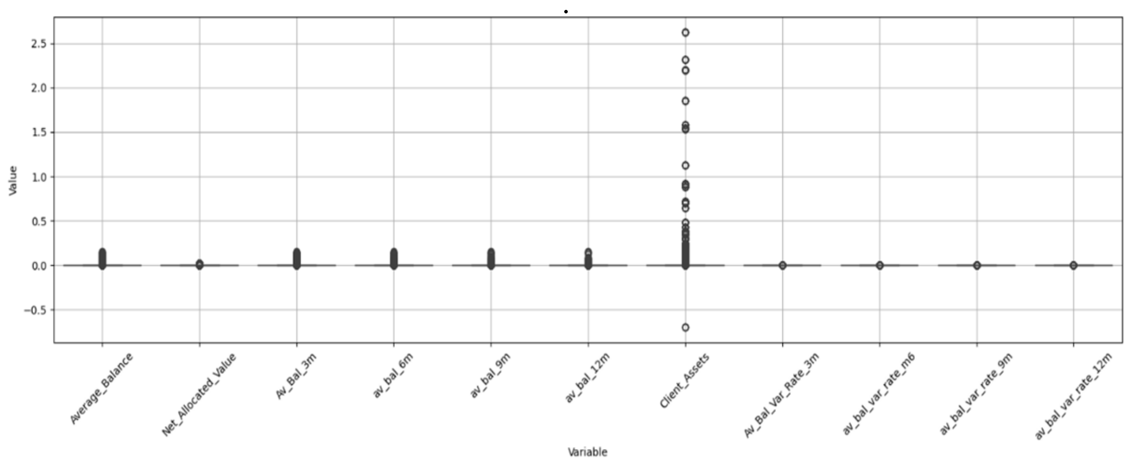

FIGURE 1: Box Plot of Continuous Variables Before Winsorization.

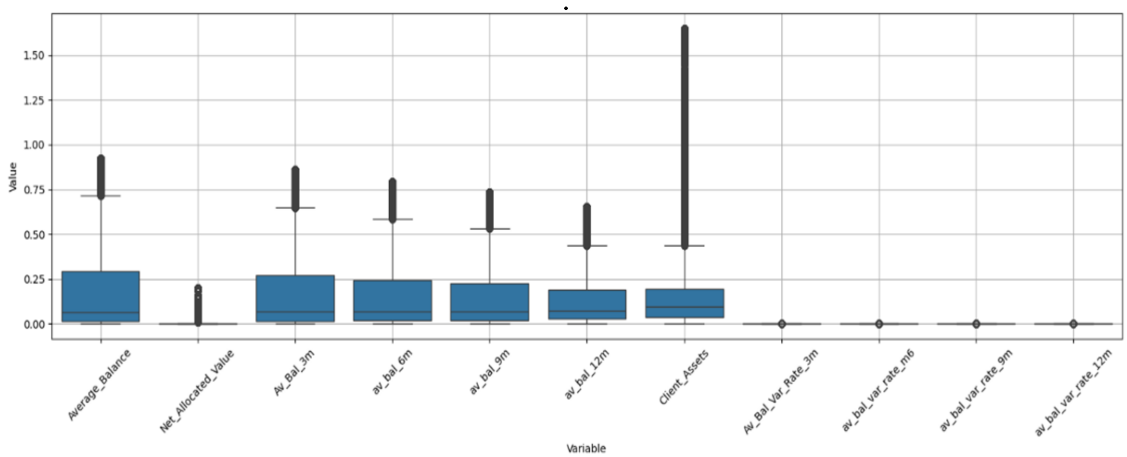

FIGURE 2: Box Plot of Continuous Variables After Winsorization.

As indicated in section 3.1, some variation rate metrics (av_bal_var_rate_3m, av_bal__var_rate_6m, av_bal_var_rate_9m, av_bal_var_rate_12m) exhibited several infinite values, specifically 147, 198, 215, and 178 instances, respectively. Null values were consistently employed to substitute infinite values, and binary variables were incorporated to indicate their initial presence. Subsequently, median imputation was employed for the remaining missing values, consistent with the approach utilized for other missing data.

The dataset was divided into training and test sets with a 70/30 ratio, utilizing distinct client identifiers (Artificial Id) to ensure an equitable evaluation of the model and prevent data leakage. This method ensures that all records for a single client appear exclusively in either the training set or the test set. This prevents information from disseminating between segments.

In order to explore the significant imbalance of the target variable (Default Flag), we employed and tested the ability of resampling approaches to enhance the model's performance for the minority class, recognizing that the quality of synthetic data imputation can be suboptimal, potentially undermining regulatory compliance. We examined three advanced techniques: SMOTE (Synthetic Minority Over-Sampling Technique), SMOTE-Tomek (a combination of oversampling and cleansing utilizing Tomek connections), and ADASYN (Adaptive Synthetic Sampling). These strategies either generate synthetic cases of the minority class or eliminate ambiguous data, hence enhancing the classifier's ability to generalize to infrequent occurrences. Subsequent sections of this study will examine the efficacy of these tactics in comparison to one another and to the natural data shape.

In addition to resampling, the CatBoost, LightGBM, and XGBoost models were assigned class weights. Class weighting assigns greater significance to the minority class (defaults) during model training, indicating that the misclassification of infrequent instances incurs a higher cost.

$$\text{class\_weight}_i = \frac{n_{\text{samples}}}{n_{\text{classes}} \times n_i} \tag{1}$$

where

$$\begin{aligned} n_{\text{samples}} &= \text{total number of samples}, \\ n_{\text{classes}} &= \text{number of classes}, \\ n_i &= \text{number of samples in class } i. \end{aligned}$$

This approach mitigates bias towards the majority class and enhances the models' efficacy in identifying high-risk candidates. The combination of resampling and class weighting solved the issue of class imbalance. They ensured that the models remained responsive to the minority class without compromising their overall efficacy.

The same models were trained with cross-validation. Cross-validation is a statistical technique that divides a dataset into many segments, or "subsets." The model is trained on one subset and evaluated on the remaining portions. This technique is repeated multiple times to ensure that each fold serves as a validation set once. The primary advantage of cross-validation is that it provides a more precise assessment of a model's performance by reducing the unpredictability associated with a single train-test split and diminishing the likelihood of overfitting.

3.3 Feature Selection

A multi-phase methodology for feature engineering and selection was employed to ensure the model was straightforward, comprehensible, and precise in result prediction.

This strategy employed advanced feature significance techniques, correlation analysis, and multicollinearity diagnostics concurrently.

3.3.1 Correlation Analysis and Heatmap

A correlation heatmap was ploted to illustrate the relationships between continuous and categorical data in pairs. This study identified several clusters of features that were highly correlated, particularly among the temporal average balance metrics. A threshold of 0.90 was established to identify and rectify significant association. Variables above this level may induce redundancy and compromise the stability of model predictions.

To mitigate this, the variables av_bal_3m, av_bal_6m, and av_bal_9m were removed. These variables were significantly correlated with both Average Balance and av_bal_12m. The decision to retain Average Balance and av_bal_12m was chosen since they encompass a broader temporal range and enhance the model's significance. The Average Balance provides a comprehensive overview of the client's financial status, whereas av_bal_12m illustrates long-term balance changes. Collectively, they provide a more comprehensive understanding of how balance evolves over time compared to measurements that focus solely on the short term.

3.3.2 Multicolinearity Control

To further address multicollinearity, the Variance Inflation Factor (VIF) was computed for all predictors.

$$\text{VIF}_j = \frac{1}{1 - R_j^2} \tag{2}$$

where

$$\begin{aligned} \text{VIF}_j &= \text{variance inflation factor for predictor } j, \\ R_j^2 &= \text{coefficient of determination of the regression of predictor } j \\ &\quad \text{on all other predictors.} \end{aligned}$$

A VIF threshold of 5 was set, a commonly accepted standard in statistical modeling to indicate problematic multicollinearity. This stands as a limitation for parametric models, as Logistic Regression. Variables exceeding this threshold can inflate the variance of coefficient estimates, reduce interpretability, and compromise model stability. As a result, Encoded_Employment_Sector and Encoded_Educ_Level were excluded from the final feature set due to their high VIF values, ensuring that retained predictors contribute unique and non-redundant information to the model.

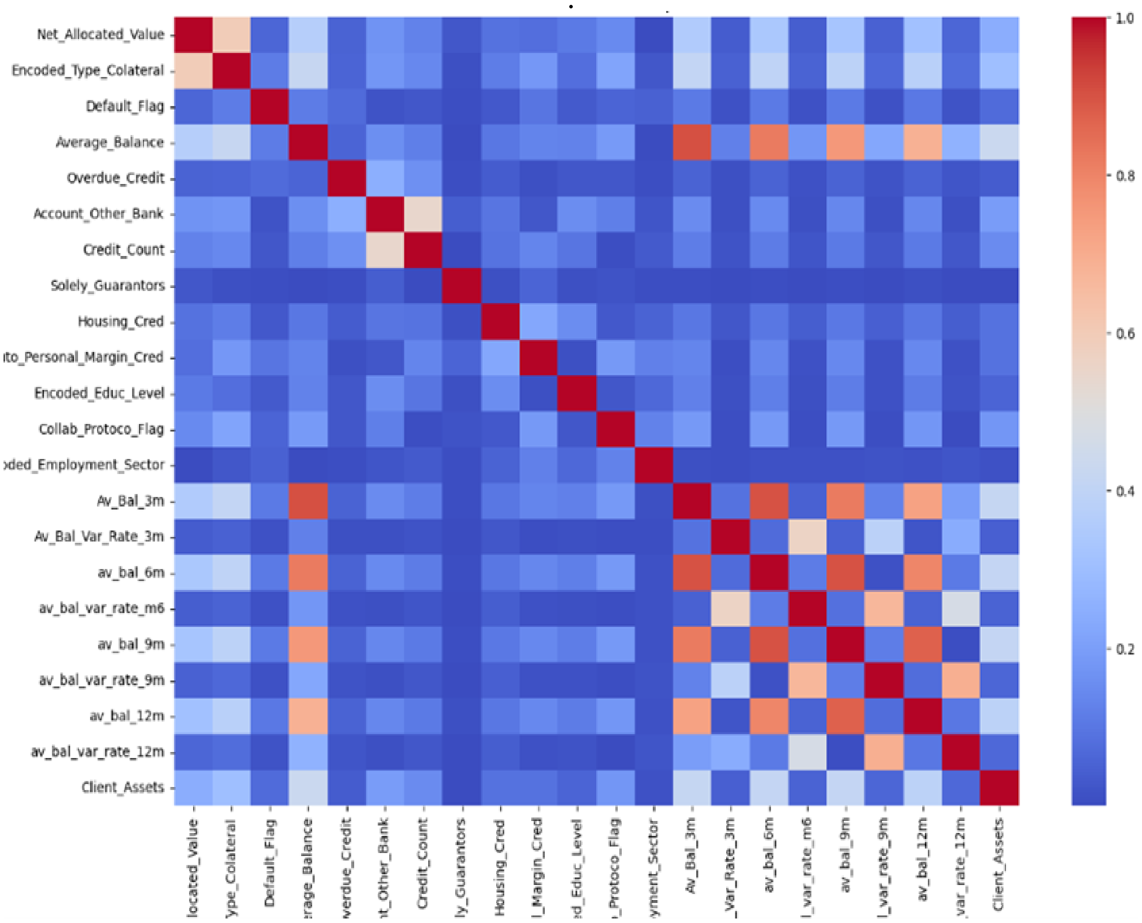

FIGURE 3: Correlation Heatmap.

3.3.3 Feature Importance and SHAP

Both model coefficients and SHAP (Shapley Additive Explanations) values for the XGBoost model were examined to assess the significance of each feature. SHAP values indicate the contribution of each feature to the model's predictions. They provide a universal metric of significance applicable to all modelling kinds. The SHAP summary plot (Figure 4) and the feature importance bar plot (Figure 5) facilitated the selection of variables for use. Features with minimal impact on SHAP, negligible benefit, or narrow value range were considered for exclusion. The subsequent factors were excluded:

- **Solely Guarantors**: Demonstrated very low SHAP impact and negligible importance across all models.
- **av_bal_var_rate_6m**: Showed low gain and contributed minimally in the SHAP analysis.
- **av_bal_var_rate_9m**: Displayed a narrow SHAP value range and low variance,

indicating limited predictive value.

- **av_bal_var_rate_3m**: Exhibited a similar pattern to the variables above, with minimal influence on model output.

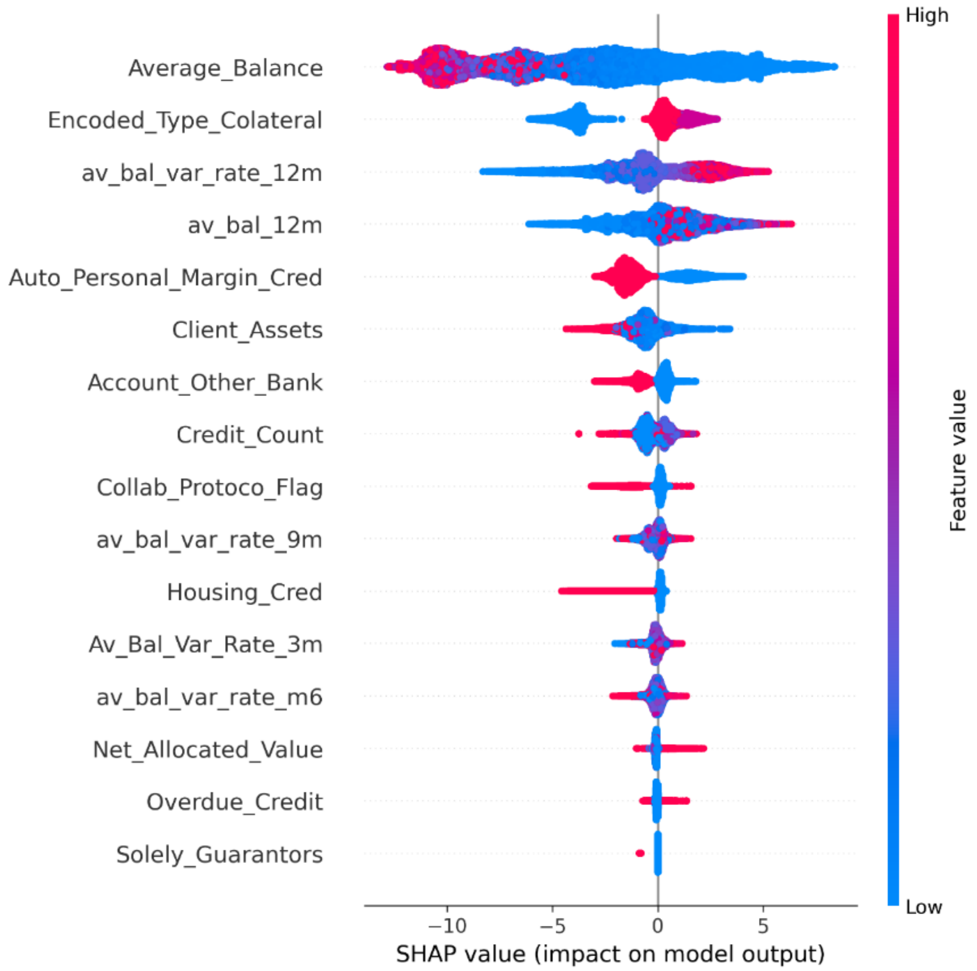

FIGURE 4: SHAP Summary plot for XGBoost model.

3.3.4 Implications of Feature Selection

The final set selected is robust, as it eliminates variables that are highly correlated or collinear, along with those of minimal significance. Retaining solely the most comprehensive and beneficial components mitigates the risk of overfitting and enhances comprehension. Employing SHAP values ensures that variable selection is grounded in the actual predictive contributions of each feature, rather than arbitrary statistical criteria.

This feature selection process establishes a solid foundation for subsequent modelling, ensuring that the final set of predictors covers the statistical challenge.

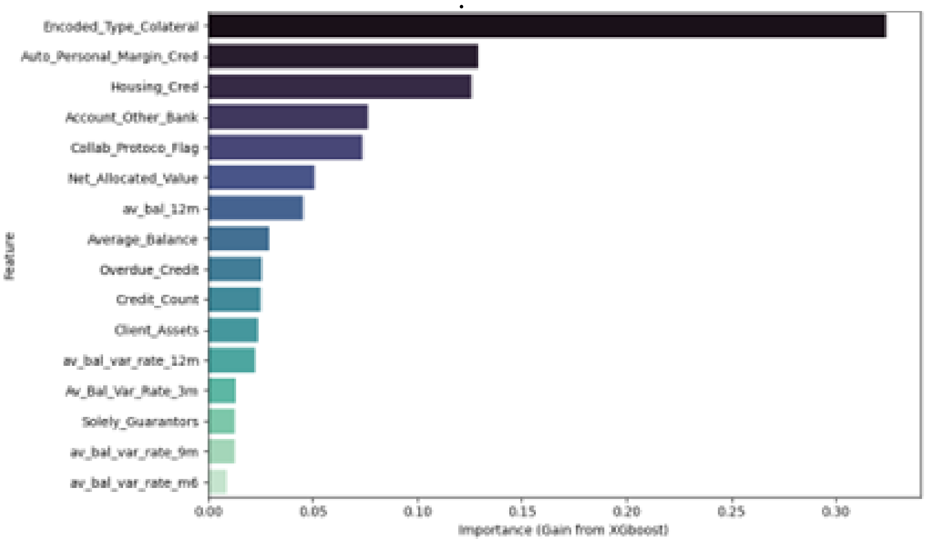

FIGURE 5: Feature Importance of XGBoost model.

3.4 Model Description and Development

3.4.1 Logistic Regression

Logistic regression is a prevalent statistical method for predicting the probability of a binary outcome based on one or more predictor variables. Logistic regression differs from linear regression as it fails to foresee continuous outcomes. Rather, it assesses the probability that a specific input is classified as either default or non-default, deeply useful in credit risk scoring. The logistic (or sigmoid) function is the fundamental component of logistic regression. It transforms a linear combination of input attributes into a probability value ranging from 0 to 1.

Mathematically, the model predicts the log-odds (logit) of the event as a linear function of the predictors:

$$\begin{aligned} \log\left(\frac{p}{1-p}\right) &= \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \\ \text{where} \quad p &= \text{probability of the outcome of interest,} \\ \alpha &= \text{intercept term,} \\ \beta_1, \ldots, \beta_k &= \text{coefficients for the predictors,} \\ x_1, \ldots, x_k &= \text{explanatory variables.} \end{aligned} \tag{3}$$

Logistic regression is a "discriminative classifier," indicating that it directly models the probability of the outcome based on the predictors. This differs from generative models, which represent the combined distribution of predictors and outcomes. The Logistic regression model provides a probability that can be thresholded, typically at 0.5, to provide a class label.

Likelihood ratio tests are frequently employed to assess the statistical significance of each predictor. The whole model fit is subsequently evaluated using metrics such as the Kolmogorov–Smirnov statistic, the area under the ROC curve (AUC), or concordance measures. Regularization techniques such as L1 (Lasso) and L2 (Ridge) can be employed to mitigate overfitting, particularly in scenarios with numerous variables or limited sample sizes.

Logistic regression is based on several key assumptions. The dependent variable must be binary, meaning it has only two possible outcomes. Secondly, the observations must be independent, meaning they do not influence each other. Third, a linear relationship must exist between the continuous predictors and the logit of the outcome. The predictors must not exhibit excessive collinearity. The technique also suggests that there are no outliers significantly influencing the outcomes, as such points could alter the model's calculations and projections.

3.4.2 Forward Stepwise Selection using BIC

Forward stepwise selection utilizing the Bayesian Information Criterion (BIC) is a systematic approach for selecting variables, facilitating the identification of a limited set of predictors for regression models, such as Logistic regression. The methodology starts with a null model including only the intercept, thereafter incorporating variables sequentially. At each level, the predictor that optimally enhances the model, as determined by the lowest BIC, is incorporated into the model. This approach continues until no other predictors can reduce the BIC.

The Bayesian Information Criterion (BIC) is a method for selecting a model that reconciles data fit with model complexity by incorporating a penalty for the number of parameters. The Bayesian Information Criterion (BIC) penalizes model complexity more severely than the Akaike Information Criterion (AIC), facilitating the comprehension and utilization of simpler models. This helps keep models from overfitting, especially when there are a lot of data points. The BIC for a given model is calculated as:

$$\begin{aligned} \mathrm{BIC} &= -2 \cdot \ln(\hat{L}) + k \cdot \ln(n) \\ \text{where} \quad \hat{L} &= \text{maximum likelihood of the model,} \\ k &= \text{number of estimated parameters,} \\ n &= \text{sample size.} \end{aligned} \tag{4}$$

The procedure for forward stepwise selection is outlined as follows: Starting with the intercept-only model, each prospective predictor is evaluated for inclusion. The option that most significantly reduces the BIC is incorporated. The approach persists, examining the remaining predictors at each stage, until no additional reduction in BIC is attainable. The result is a collection of interrelated models. The latest model is superior as it achieves an optimal equilibrium between clarity and comprehensive explanation, as per the BIC criterion.

This strategy is highly beneficial in practical scenarios where comprehension and generalization are crucial, as demonstrated in this study. It is essential to acknowledge that stepwise approaches, such as those utilizing BIC, may exhibit sensitivity to the order of variable inclusion and may not consistently identify the true underlying model.

3.4.3 Catboost

CatBoost is an innovative gradient boosting technique designed to effectively handle both numerical and categorical variables in supervised learning contexts. CatBoost is an ensemble methodology that generates a sequence of decision trees. Every tree is instructed to rectify the errors of the prior ensemble. This enables the model to identify intricate, non-linear relationships within the data. This iterative boosting approach is highly effective for both classification and regression problems.

One distinguishing feature of CatBoost is its natural capability to manage categorical variables. Both logistic regression and stepwise selection methods require the explicit encoding of categorical variables, such as by one-hot encoding or ordinal transformation. Conversely, CatBoost autonomously transforms categorical features utilizing ordered target statistics. This approach use permutations and meticulous calculations to reduce target leakage and overfitting, enabling the algorithm to extract greater insights from categorical data with minimal preprocessing.

CatBoost employs symmetric trees, indicating that all nodes at a specific depth utilize the same rule for partitioning. This framework enhances prediction speed and reliability, reduces overfitting, and increases model robustness, particularly when the dataset comprises diverse feature types. Furthermore, CatBoost employs ordered boosting, which rectifies prediction shift and facilitates generalization to novel data, particularly in the presence of diverse variable types or limited sample sizes.

From a performance perspective, CatBoost has demonstrated superior predictive accuracy and robustness compared to traditional models, anticipating that CatBoost will represent a significant advancement in methodology relative to traditional techniques such as Logistic regression and stepwise selection. It is particularly effective for managing complex, heterogeneous datasets typical in real-world applications, as it can inherently accommodate categorical variables, identify non-linear connections, and generate precise forecasts with minimal preparation.

3.4.4 LightGBM

LightGBM (Light Gradient Boosting Machine) is an advanced iteration of the gradient boosting approach designed for enhanced speed and precision, particularly suited for large and complex datasets, which is the focus of this study. LightGBM generates an ensemble of decision trees sequentially, similar to previous boosting techniques. Every new tree is cultivated to mitigate the errors residual from the preceding ensemble. This iterative approach enables the LightGBM model to capture complex, non-linear relationships and interactions among features. This differs from logistic regression and stepwise selection, which are limited to predicting linear combinations of input variables.

The leaf-wise tree development method is among the most significant innovations introduced by LightGBM. Level-wise growth is a prevalent technique employed in traditional boosting algorithms, encompassing both standard implementations and in CatBoost. In this approach, all nodes at a specific level are fully explored prior to progressing deeper. However, LightGBM constructs trees by consistently partitioning the leaf that most effectively reduces the objective loss function. This approach to analyzing leaves can enhance the depth and expressiveness of trees, hence assisting the model in identifying more nuanced patterns within the data. This often enhances the model's accuracy compared to linear models and various boosting frameworks; however, it also increases the risk of overfitting, necessitating rigorous regularization through parameters such as maximum tree depth or the number of leaves.

The histogram-based decision tree learning in LightGBM distinguishes it from conventional statistical models and alternative boosting methods. Rather than examining every possible split point for continuous variables, LightGBM categorizes feature values into a predetermined number of bins. This significantly reduces memory usage and processing time. This new feature enhances the model's efficacy for extensive datasets, when conventional boosting or regression models may struggle to manage the volume of data.

Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) are two innovative settings in LightGBM. GOSS utilizes computational capacity to concentrate on data points exhibiting the steepest gradients, which are the most challenging to predict. It additionally samples simpler instances at random, thereby accelerating training without compromising prediction accuracy. EFB amalgamates characteristics that cannot coexist simultaneously, hence diminishing the effective count of features. This is particularly advantageous for sparse, high-dimensional datasets. Logistic regression lacks these approaches, thus it does not address computational scalability or feature sparsity.

LightGBM autonomously addresses missing and categorical data by determining the optimal method for managing missing values at each split. For categorical variables, LightGBM allows users to designate which features are categorical and employs a distinct method to identify splits. Conversely, LightGBM's approach may require supplementary user input or explicit preparation, as it does not inherently encode and manage categorical information in the manner that CatBoost does. CatBoost use sophisticated target statistics to mitigate overfitting.

In terms of comprehensibility, LightGBM is less transparent than Logistic regression due to its nature as a tree-based ensemble method. The latter provides explicit coefficients for each predictor, facilitating comprehension and inference. However, LightGBM excels in predictive accuracy and adaptability, particularly as the volume and intricacy of the dataset increase, despite a reduction in interpretability.

3.4.5 XGBoost

XGBoost (Extreme Gradient Boosting) is an exceptionally efficient and scalable variant of the gradient boosting method, renowned for its superior performance compared to other algorithms when applied to structured data and real-world scenarios. Fundamentally, XGBoost develops an ensemble of decision trees sequentially, training each subsequent tree to mitigate the errors remaining from the previously constructed ensemble. This iterative boosting technique enables XGBoost to identify complex, non-linear relationships and interactions among features.

Second-order optimization is a major innovation in XGBoost. Conventional gradient boosting methods utilize solely the first derivative (gradient) of the loss function to adjust model parameters. Conversely, XGBoost employs both the first and second derivatives (the Hessian) during its optimization procedure. This enhances the precision and reliability of updates, hence accelerating convergence and improving the accuracy of forecasts. Conversely, Logistic regression and stepwise models employ maximum likelihood estimation and do not improve with more intricate optimization techniques.

Additionally, XGBoost incorporates many regularization techniques into its objective function, including L1 (lasso) and L2 (ridge) penalties. These regularization parameters help maintain model simplicity and mitigate overfitting, a potential issue with robust ensemble methods. Regularization can also be applied in logistic regression, however it is not inherently incorporated into stepwise selection. Tree-based ensembles such as XGBoost, CatBoost, and LightGBM typically require more sophisticated regularization due to their capacity to manage larger datasets.

From a computational perspective, XGBoost is designed for speed and scalability. It employs an innovative method for identifying splits that accounts for sparsity and autonomously manages missing information. This renders it ideal for extensive, high-dimensional datasets. This resembles the internal handling of missing data by LightGBM and CatBoost.

XGBoost provides parallel and distributed processing, enabling rapid model training even with extensive datasets. The block structure for data storage and processing enhances memory utilization and performance. LightGBM similarly emphasizes rapidity and scalability via histogram-based learning and leaf-wise expansion. Nonetheless, XGBoost has gained popularity in data science competitions and commercial applications due to its capability for parallel execution and cache optimization.

XGBoost requires categorical characteristics to be manually inputted for categorical data.

Another distinction is the level of comprehensibility. Similar to other tree-based ensembles, XGBoost is more complex to comprehend. However, XGBoost provides tools for assessing feature importance and visualizing decision trees, which partially addresses this issue.

3.4.6 Model Calibration

Model calibration is crucial for evaluating credit risk, since it ensures that the probability estimates generated by predictive models are precise and accurately represent the actual likelihood of default occurrences. Well-calibrated probabilities are essential for regulatory capital calculations, pricing decisions, and risk-adjusted profitability assessments in financial risk modeling. The calibration technique addresses the significant distinction between discriminative capability and probabilistic precision. This occurs because models can effectively rank while producing inaccurate probability estimations.

Calibration relies on the principle that expected probabilities must align with observed frequencies across various probability ranges. This condition is particularly crucial in credit risk applications, as persistent underestimation or overestimation of default probability can result in poor lending decisions and potential legal infractions.

Due to the substantial size of the dataset included in this study (exceeding 100,000 observations), isotonic regression was employed as the primary calibration approach. This non-parametric method exhibits greater flexibility than parametric methods, as it may accommodate any monotonic connection between uncalibrated model outputs and actual probabilities.

Isotonic regression calibration is based on the assumption that the relationship between predicted scores and actual probabilities is monotonic (non-decreasing). The method identifies a piecewise constant function that maintains the sequence of predictions while enhancing calibration accuracy. This approach is most effective when:

- Large datasets provide sufficient statistical power for reliable non-parametric estimation;
- Non-linear miscalibration patterns are observed in preliminary analysis;

When employing calibration techniques, it was essential to consider the specific compatibility concerns inherent to each model. The LightGBM and CatBoost models performed exceptionally well using *scikit-learn*'s calibration framework, enabling the use of isotonic regression. This compatibility facilitated cross-validation during calibration and enhanced memory management for large datasets.

The Brier score served as the primary quantitative metric for evaluating calibration quality, defined as:

$$\text{Brier Score} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2 \tag{5}$$

where

$$\begin{aligned} p_i &= \text{predicted probability for observation } i, \\ y_i &= \text{true binary outcome for observation } i, \\ N &= \text{total number of observations.} \end{aligned}$$

The Brier score can be decomposed into three components: reliability (calibration error), resolution (discriminative ability), and uncertainty (inherent randomness in the data). This decomposition provides valuable insights into the specific sources of prediction errors and guides calibration improvement efforts.

The calibration process required careful integration with cross-validation procedures to ensure robust performance estimates. Calibration parameters were estimated using data independent of both the model training set and the final evaluation set, necessitating a three-way data split to prevent overfitting and ensure generalization of calibration improvements.

The selection of isotonic regression for large datasets balanced theoretical optimality with computational efficiency requirements. The non-parametric nature of isotonic regression enables accurate modeling of complex calibration relationships while maintaining reasonable computational overhead for quality maximization while ensuring robust, deployable solutions suitable for such emerging credit risk environments.

3.4.7 Evaluation Metrics

To effectively analyze classification models for credit risk evaluation, a comprehensive approach is required that addresses the challenges posed by unbalanced datasets, where default instances constitute a mere fraction of the observations. This strategy employs various assessment metrics, each evaluating a distinct aspect of anticipated performance. Also employing threshold optimization methods to enhance the model's efficacy.

In credit risk modeling, the confusion matrix is the key component among all metrics utilized for categorization evaluation. This matrix sets the parameters of model prediction into four groups: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). In this instance:

- TP: correctly identified defaulters,
- FP: non-defaulters incorrectly flagged as high-risk,
- TN: correctly identified non-defaulters,
- FN: defaulters incorrectly classified as low-risk.

3.4.8 Fundamental Classification Metrics

Precision

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

Working on minimizing false alarms, Precision measures how many of the positively predicted cases (defaults) are actual defaulters. It is critical when false positives are costly.

Recall

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

The metric to capture all defaults, Recall focuses on the model's ability to detect all actual defaulters. It is essential for minimizing the financial risks associated with missed defaults.

F1-Score

$$F_1(Score) = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

The F1-score penalizes extreme values of accuracy and recall, rendering it an optimal selection for threshold optimization in credit risk, since it embodies the model's balance, relying in both precision and recall simultaneously.

ROC-AUC and PR-AUC will be used as advanced evaluation metrics instead of the baseline metric.

The Receiver Operating Characteristic (ROC) curve illustrates the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) across various classification thresholds. The Area Under the Curve (ROC-AUC) is a threshold-independent metric that measures a model's ability to distinguish between positive and negative classes. It ranges from 0.0 to 1.0, where:

- $[0,0;0,5[$: indicates a model performing worse than random guessing;
- $= 0,5$: indicates a model with no discriminative ability (equivalent to random guessing);
- $]0,5;1,0[$: indicates a model that can distinguish classes to some degree, with higher values reflecting better performance;
- $= 1$: perfect discrimination, perfectly separating positive and negative classes;

As TPR and FPR are ratios particular to each class, they remain effective even in scenarios where one class predominates over the other.

The Precision-Recall (PR) curve illustrates the performance of the minority class. The PR-AUC is highly sensitive to class imbalance, which is advantageous for credit risk due to the infrequency of defaults. It provides a clear assessment of the model's efficacy in identifying non-payers.

ROC-AUC is effective for comparing models across different datasets and time periods, whereas PR-AUC demonstrates a model's performance in real-world scenarios with imbalanced data. We will examine both measures concurrently due to their synergistic effectiveness.

The selection of the classification threshold (θ) significantly impacts performance, particularly when the datasets are imbalanced. A default threshold of 0.5 frequently yields suboptimal outcomes. The F1-score is a balanced metric that illustrates the trade-off between false positives and false negatives, rendering it an excellent target for threshold adjustment. To determine the optimal threshold θ^*, we seek the value that maximizes the F1-score.

$$\theta^* = \arg\max_{\theta} F_1(\theta) \tag{9}$$

where

$$\begin{aligned} \theta^* &= \text{optimal threshold,} \\ F_1(\theta) &= \text{F1-score evaluated at threshold } \theta. \end{aligned}$$

The optimization process typically involves: Grid Search Evaluation, testing thresholds between 0.01 and 0.99.

3.5 Results and Discussion

This chapter presents the systematic evaluation of the presented models using the various resampling techniques introduced.

3.5.1 Overview of Model Performance

Table II illustrates that the models exhibit significantly distinct operational features. In the absence of resampling, CatBoost exhibited the superior classification performance, achieving an F1-score of 0.34 and demonstrating a commendable equilibrium between precision (0.35) and recall (0.34) at a threshold of 0.1441. The low threshold setting facilitated the identification of genuine defaulters while maintaining an acceptable false positive rate, representing an optimal approach for risk-based credit allocation.

The combination of LightGBM and ADASYN resampling achieved the maximum recall (0.36), which is advantageous for businesses aiming to minimize the occurrence of missed defaults. Nonetheless, the enhancement in recall was counterbalanced by a notable decline in precision (0.20), indicating that severe oversampling of the minority class typically results in an increase in the proportion of false positives. XGBoost with SMOTE prioritized recall (0.38) at the cost of accuracy (0.22), exemplifying the typical trade-off between recall and precision. Conversely, Logistic regression and its stepwise variant yielded lower F1-scores ($\leq$ 0.22); nonetheless, they provide clear and comprehensible decision boundaries that are advantageous in regulated contexts.

Table III presents the ROC-AUC and PR-AUC metrics, which provide additional insights into the model's ability to differentiate across classes, particularly in the presence of significant class imbalance. LightGBM performed optimally on the unprocessed dataset,

TABLE II: Classification Metrics and Thresholds for All Models

Model	F1-Score	Precision	Recall	Threshold Used
Logistic Regression	0.17	0.12	0.27	0.7740
Logistic Regression (SMOTE)	0.21	0.23	0.19	0.8566
Logistic Regression (SMOTETomek)	0.22	0.26	0.19	0.8706
Logistic Regression (ADASYN)	0.20	0.26	0.17	0.8712
Forward Stepwise Selection	0.19	0.14	0.29	0.0929
Forward Stepwise (SMOTE)	0.21	0.35	0.15	0.8953
Forward Stepwise (SMOTETomek)	0.21	0.34	0.15	0.9138
Forward Stepwise (ADASYN)	0.20	0.28	0.16	0.8893
CatBoost	0.34	0.35	0.34	0.1441
CatBoost (SMOTE)	0.29	0.31	0.28	0.7621
CatBoost (SMOTETomek)	0.29	0.33	0.26	0.8202
CatBoost (ADASYN)	0.25	0.29	0.22	0.8076
LightGBM	0.31	0.36	0.27	0.1676
LightGBM (SMOTE)	0.31	0.38	0.26	0.8272
LightGBM (SMOTETomek)	0.29	0.30	0.29	0.7816
LightGBM (ADASYN)	0.26	0.20	0.36	0.5312
XGBoost	0.25	0.23	0.28	0.5484
XGBoost (SMOTE)	0.28	0.22	0.38	0.5061
XGBoost (SMOTETomek)	0.28	0.30	0.27	0.7862
XGBoost (ADASYN)	0.27	0.30	0.25	0.7471

TABLE III: Discrimination Metrics for All Models

Model	**ROC-AUC**	**PR-AUC**
Logistic Regression	0.774	0.0862
Logistic Regression (SMOTE)	0.747	0.1229
Logistic Regression (SMOTETomek)	0.749	0.1291
Logistic Regression (ADASYN)	0.722	0.1285
Forward Stepwise Selection	0.779	0.0932
Forward Stepwise (SMOTE)	0.751	0.1050
Forward Stepwise (SMOTETomek)	0.755	0.1066
Forward Stepwise (ADASYN)	0.716	0.1323
CatBoost	0.847	0.2414
CatBoost (SMOTE)	0.835	0.1923
CatBoost (SMOTETomek)	0.833	0.1892
CatBoost (ADASYN)	0.825	0.1634
LightGBM	0.850	0.2270
LightGBM (SMOTE)	0.848	0.1986
LightGBM (SMOTETomek)	0.845	0.1895
LightGBM (ADASYN)	0.839	0.1711
XGBoost	0.825	0.1961
XGBoost (SMOTE)	0.846	0.1940
XGBoost (SMOTETomek)	0.845	0.1868
XGBoost (ADASYN)	0.841	0.1855

TABLE IV: Performance Metrics for Calibrated Models

Model	F1-Score	Precision	Recall	ROC-AUC	PR-AUC	Brier Score
CatBoost	0.33	0.40	0.28	0.856	0.2358	0.0281
LightGBM	0.32	0.30	0.34	0.865	0.2402	0.0280
XGBoost	0.27	0.25	0.28	0.814	0.1772	0.0514

achieving a ROC-AUC value of 0.850 and a PR-AUC value of 0.2270. This indicates a strong ability to differentiate between individuals who default and those who do not.

CatBoost exhibits the highest PR-AUC at 0.2414, closely succeeded by ROC-AUC at 0.847. This demonstrates its efficacy in identifying actual positive instances, even in the context of infrequent events. Logistic regression with SMOTETomek and the stepwise variant, in conjunction with ADASYN, yielded the highest PR-AUC values among linear methodologies. In this scenario of imbalance, PR-AUC is crucial as it is highly sensitive to the system's efficacy in retrieving the rare, high-stakes positive class.

3.5.2 Calibration Analysis

The Brier Score, derived from isotonic regression, assesses the reliability of probabilistic forecasts. The utilization of credit rating algorithms in practical financial contexts is of paramount significance. The calibration findings, presented in Table IV, indicate that LightGBM achieved the lowest Brier Score (0.0280) and the highest ROC-AUC (0.865) post-calibration. This indicates that it excelled in both probability estimation and total discrimination.

CatBoost demonstrated consistent performance, achieving a Brier Score of 0.0281, closely approaching the optimal score, and a PR-AUC of 0.2358, which was the highest recorded. This demonstrates its efficiency. XGBoost exhibited a significantly elevated Brier Score (0.0514) and a diminished PR-AUC (0.1772), indicating that its probabilistic outputs are less dependable compared to its ensemble counterparts, despite improved performance post-calibration.

3.5.3 Effects of Resampling Strategies

The base model architecture significantly influenced the effect of resampling strategies. The SMOTETomek method provided the optimal balance between precision and recall, particularly for linear models. ADASYN increased the significance of recall; yet, it simultaneously rendered precision less stable, a consequence of attempting to accommodate an excessive number of synthetic minority instances.

It is noteworthy that both CatBoost and LightGBM, in their unaltered states without resampling, outperformed all resampling-enhanced variants. The improved management of class imbalance inherent in contemporary gradient boosting methods demonstrates their robustness. This also prompts critical inquiries regarding the timing and methodology of employing synthetic data augmentation, particularly for high-capacity models operating in low-quality data contexts.

3.5.4 Strategic Recommendations for the Angolan Banking Sector

It is essential to select the appropriate model for each institution in accordance with its strategic objectives, regulatory mandates, and risk appetite, informed by the comprehensive results of the categorization, discrimination, and calibration analyses.

Logistic regression, particularly when integrated with SMOTETomek resampling to address class imbalance, remains highly advisable in contexts where transparency and regulatory elucidation are paramount.

To enhance risk prediction and automated scoring, and to achieve optimal performance metrics such as F1-score, PR-AUC, and ROC-AUC, your organization should utilize either uncalibrated CatBoost or calibrated LightGBM models. These models possess robust predictive capabilities and precise probability assessments, aiding individuals in making decisions in critical scenarios.

LightGBM using ADASYN resampling is the optimal selection when the primary objective of the institution is to minimize missed defaults. This is due to its elevated recall rate, indicating its capacity to identify numerous risk situations, but at the expense of less precision. XGBoost models, whether incorporating SMOTETomek or not, demonstrate optimal accuracy in scenarios where the primary objective is to minimize false positives and prevent costly credit denials.

It is crucial to recognize that SHAP values are particularly applicable to ensemble models such as CatBoost and LightGBM. Employing SHAP enhances clarity by providing interpretability levels akin to those of logistic regression. This solution significantly narrows the disparity between the opaque characteristics of sophisticated ensemble methods and the necessity for transparency in regulated financial services. This enables financial institutions to utilize optimal predictive models while maintaining comprehensibility.

4 Conclusion

This empirical study has elucidated the equilibrium between understanding credit risk models and their effectiveness in generating precise predictions inside the Angolan mortgage market. Although classic Logistic regression models remain valuable due to their simplicity and auditability, contemporary ML, as tree-based ensemble learning techniques, particularly CatBoost and LightGBM, have demonstrably surpassed them in both discriminative ability and the provision of well-calibrated probability estimates.

The application of isotonic regression for post hoc calibration has significantly enhanced the dependability of probability projections, which is crucial. This is particularly crucial for banks and other financial institutions that must adhere to rigorous regulations, as they require precise and well-calibrated risk evaluations to make sound credit choices. Model selection strategies must align with the institution's operational objectives, legal responsibilities, and comprehensive risk management approach. This congruence enables the development of robust, effective and reliable credit risk assessment systems. These platforms can enhance access to credit for a greater number of individuals and increases emerging market's resilience.

The Angolan financial sector consistently grapples with poor data quality, which complicates the contextualization of this research. The absence of comprehensive and consistent credit histories, the recurrent lack of values, the presence of noisy records, and the significant number of outliers have impeded the development and evaluation of predictive models. Factors contributing to these data gaps include the insufficient number of consolidated credit bureaus and the varying regulations among banks on data collection and reporting.

Poor data quality has numerous consequences: it increases the risk of data-driven biases, complicates the model's ability to represent population variety, and can lead to instability and overfitting in models. This study required a robust preprocessing approach to address these issues, encompassing meticulous missing value imputation, outlier control, modification of problematic measures, and consistent utilization of indicator variables to inform models of missing original data. To enhance the model's generalization and reliability, they employed sophisticated calibration and resampling techniques to address the inherent volatility and imbalance of the target dataset.

Despite these challenges, research indicates that significant enhancements in predictions can be achieved even in difficult circumstances by meticulous data handling and the application of advanced ML techniques. However, Angola's persistent low-quality data hampers the comprehensive utilization of credit risk analytics. For future Angolan advancement, it is imperative that all sectors enhance their data infrastructure, reporting mechanisms, and incentives for lenders to maintain comprehensive and precise records.

Ensemble methodologies that integrate the strengths of CatBoost, LightGBM, and Logistic regression via model stacking appear to be a promising direction for the future.

Such models are optimal for ensuring that resilience and performance remain consistent when managing extensive and diverse loan portfolios. They also compensate for certain issues arising from the utilization of flawed data.

This thesis elucidates the necessity of a deep change in credit risk modeling frameworks not just to address technical issues such as data imbalance and model calibration but also to account for real-world factors such insufficient data quality. By exercising caution in their methodologies and employing appropriate models, Angola and other emerging economies could achieve significant advancements in reliable, beneficial, and accessible credit risk assessment.

References

Andrianova, S., Baltagi, B., Demetriades, P., & Fielding, D. (2014). Why do African banks lend so little? *Oxford Bulletin of Economics and Statistics*, 77(3):339–359.

Song, P. (2025). How does logistic regression handle non-linear relationships? *ML Journey*.

Xia, Y., Han, Z., Li, Y., & He, L. (2025). Credit scoring model for fintech lending: An integration of large language models and FocalPoly loss. *International Journal of Forecasting*, 41(3), 894–919.

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: Gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30.

Fonseca, P. G., & Lopes, H. D. (2017). Calibration of machine learning classifiers for probability of default modelling. *arXiv preprint arXiv:1710.09845*.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Chen, Y., Calabrese, R., & Martin-Barragán, B. (2024). Interpretable machine learning for imbalanced credit scoring datasets. *European Journal of Operational Research*, 312(1), 357–372.

Gai, Y. (2024). Research on the application of alternative data in credit risk management. *Highlights in Business, Economics and Management*, 40, 1156.

Bastos, J. A. (2022). Predicting credit scores with boosted decision trees. *Forecasting*, 4(4), 925–935.

Zedda, S. (2024). Credit scoring: Does XGBoost outperform logistic regression? A test on Italian SMEs. *Research in International Business and Finance*, 70, Part B, 102397.

Bastos, J. A., & Matos, S. M. (2022). Explainable models of credit losses. *European Journal of Operational Research*, 301(1), 386–394.

Wang, H., & Cheng, L. (2021). CatBoost model with synthetic features in application to loan risk assessment of small businesses. *arXiv preprint arXiv:2106.07954*.

Choueiry, G. (2024). What is an acceptable value for VIF? *Quantifying Health Blog*.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. New York: John Wiley & Sons.

Basel Committee on Banking Supervision (2000). Credit risk modelling: Current practices and applications. *Bank for International Settlements*.

Bluhm, C., Overbeck, L., & Wagner, C. (2010). *Introduction to Credit Risk Modeling* (2nd ed.). New York: Chapman and Hall/CRC.

Călin, A. C., & Popovici, O. C. (2014). Modeling credit risk through credit scoring. *Internal Auditing & Risk Management*, 9(2[34]):99–109.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance*, 29(2):449–470.

Wang, Z. (2024). Artificial intelligence and machine learning in credit risk assessment: Enhancing accuracy and ensuring fairness. *Open Journal of Social Sciences*, 12(11).

MITSDE (2024). Data imputation techniques: Handling missing data in machine learning. *MITSDE Blog*.