LISBON
SCHOOL OF
ECONOMICS &
MANAGEMENT
UNIVERSIDADE DE LISBOA

MASTER

DATA ANALYTICS FOR BUSINESS

MASTER´S FINAL WORK

DISSERTATION

MODELS FOR THE PROBABILITY OF MORTGAGE DEFAULTS

MARIANA DOS SANTOS RIBEIRO DIAS

SUPERVISION:

PROF. JOÃO AFONSO BASTOS

FEBRUARY - 2025

Glossary

AIW Average Interval Width. i, 25, 26, 31–33

AUC Area Under the Curve. i, 5, 11, 16, 17, 21, 23, 28, 29

CV Cross-Validation. i, 15, 24

EDA Exploratory Data Analysis. i, 8

KNN K-Nearest Neighbors. i, 8

LAC Least Ambiguous set-valued Classifier. i, 24

LightGBM Light Gradient-Boosting Machine. i, ii, 2, 4, 19, 20, 29, 30, 33

MAPIE Model-Agnostic Prediction Interval Estimation. i, ii, 1–3, 6, 22–25, 28, 31–34

ML Machine Learning. i, 2, 4

PR Precision-Recall. i, 5, 11, 21, 23, 28, 29

RFE Recursive Feature Elimination. i, 13, 16, 17

ROC Receiver Operating Characteristic. i, 5, 11, 16, 17, 23, 28, 29

SHAP SHapley Additive exPlanations. i, 5, 13–17

SMOTE Synthetic Minority Oversampling Technique. i, 21, 23, 28–33

VIF Variance Inflation Factor. i, 12, 13

XGBoost Extreme Gradient Boosting. i, ii, 2–5, 13, 15–17, 19, 20, 28–30, 33

ABSTRACT, KEYWORDS, AND CODES

Mortgage default prediction is a critical task for financial institutions, where accurately identifying high-risk borrowers is essential for mitigating financial losses and ensuring responsible lending practices. Traditional credit scoring models, such as logistic regression, are widely used but often fail to capture complex patterns in borrower behaviour, especially when the data is highly skewed.

This thesis applies Logistic Regression, Random Forest, Extreme Gradient Boosting (XGBoost) and Light Gradient-Boosting Machine (LightGBM) to mortgage default data, using several feature selection techniques and data imbalance strategies. Calibration is also applied with both Platt Scaling and Isotonic Regression, and their performances are evaluated.

In addition, the behaviour of Model-Agnostic Prediction Interval Estimation (MAPIE) in the context of mortgage default prediction is investigated. By leveraging MAPIE's conformal prediction framework, this study assesses its ability to provide robust uncertainty estimates and reliable predictive intervals for default classification.

The results obtained demonstrate that, for this dataset, boosting models, particularly XGBoost, outperform Logistic Regression in mortgage default prediction. Addressing class imbalance through hybrid resampling techniques was the most beneficial for the Random Forest model, while boosting methods hand class imbalance better by using built-in parameters. Isotonic Regression worked well for tree-based algorithms, while Platt Scaling was better for Logistic Regression. When using Model-Agnostic Prediction Interval Estimation (MAPIE), balancing coverage and interval width was a challenge, making it necessary to use another metric that took both into account: Exact Match Rate. These findings highlight the importance of combining advanced machine learning techniques with calibration and uncertainty quantification to improve risk assessment in financial institutions, offering a more data-driven and reliable approach to credit decision-making.

KEYWORDS: Mortgage Default Prediction; Credit Scoring; Imbalanced Dataset; Binary Classification Problems; Calibration; MAPIE.

CODES: C52; C53; C55; G21; G33; M21.

TABLE OF CONTENTS

List of Figures

LIST OF TABLES

ACKNOWLEDGEMENTS

I wish to thank Professor João Bastos for his encouragement, advice, and guidance.

In addition, I am grateful to my colleagues Yuliya Pak and Carolina Conchinha, with whom I always teamed up with throughout the masters.

I would also like to thank my work colleagues at Closer, who I have learned so much with.

A word of appreciation is deserved for my partner, Rodrigo, for his daily support and encouragement throughout the whole masters.

Finally, I am so thankful to my family and friends for their patience and their support while I pursued this project. In particular, I would like to thank my friend Raquel and my brother João, who were consistently available to give me their wise words and take a look at my work whenever I came to them for advice, and my parents, Clarisse and Jaime, who have always prioritised my education.

MODELS FOR THE PROBABILITY OF MORTGAGE DEAFULTS

By Mariana Dias

This thesis explores machine learning models for mortgage default prediction, comparing Logistic Regression, Random Forest, XGBoost, and LightGBM. Feature selection techniques, class imbalance strategies, and probability calibration (Platt Scaling and Isotonic Regression) are applied. Additionally, MAPIE is evaluated for uncertainty quantification. Findings show that boosting models, particularly XGBoost, outperform Logistic Regression, while Random Forest benefits most from hybrid resampling. Isotonic Regression improves calibration for tree-based models. Balancing MAPIE's coverage and interval can be challenging, so another metric was used - Exact Match Rate.

1 INTRODUCTION

The rapid growth of data in financial institutions has transformed credit risk assessment, making it a cornerstone of modern banking and lending processes. Credit scoring, the method used to evaluate the likelihood of default by borrowers, plays a pivotal role in mitigating financial risks and aiding in lending decisions. Credit scoring can be a quite complex problem, therefore it could benefit from recent developments in the Machine Learning (ML) subject. However, the financial sector tends to be more conservative in their technological and analytical approaches, being somewhat resistant to change and to the adoption of recent innovations. Traditional credit scoring models, such as Logistic Regression, are still widely used, but struggle with complex relationships in the data (Zedda 2024). This can become an even worse issue when datasets are large, heterogeneous, and imbalanced.

Imbalanced datasets are a recurring challenge in real-life data analysis and predictions, namely in the security and the financial sectors (He & Garcia 2009). Credit scoring, is no exception, as default cases typically constitute a small fraction of the observations. This imbalance can bias predictive models toward the majority class (no default), leading to suboptimal performance in detecting defaults. Given the high financial stakes of misclassification in credit scoring, it is essential to find ways that will mitigate the imbalance issue.

In this paper, four different model implementations are described, as well as their results, for the purpose of mortgage default prediction: Logistic Regression, for its widespread use in real-life situations and its interpretability, Random Forest, as it has been noted to perform well in such scenario (Li & Wu 2024), and boosting methods, namely XGBoost and LightGBM, as they almost always are able to surpass Logistic Regression (Zedda 2024). In addition, strategies to handle data imbalance are explored, as this is a real concern in practical applications of these models.

Another issue with many of these ML models lies in their poor calibration, which leads to unreliable probability estimates (Gupta et al. 2022). This study applies probability calibration techniques, namely Platt Scaling and Isotonic Regression, to help refine predicted probabilities, making them more interpretable and trustworthy for risk assessment.

While probability calibration improves the reliability of individual predictions, it does not quantify uncertainty. To address this, the last step of this paper incorporates MAPIE, a conformal prediction framework that provides uncertainty estimates, using confidence levels, as risk assessment requires high confidence in predictions. In addition, it also provides interpretability. Most applications of MAPIE focus on regression tasks, making this extension to binary classification in mortgage default prediction an interesting and novel research direction.

Moreover, while existing studies have explored machine learning techniques like XGBoost and neural networks for credit scoring, few have focused on integrating conformal prediction methods, particularly MAPIE, in such contexts. This thesis seeks to fill this gap by evaluating MAPIE's performance in predicting credit defaults using an imbalanced dataset. By combining advanced predictive techniques with robust uncertainty quantification, this thesis aims to contribute to the growing body of knowledge on credit risk modelling.

The remainder of this study is structured as follows. Chapter 2 provides a detailed review of the existing literature on credit scoring, imbalanced datasets, calibration, and uncertainty quantification techniques, with an emphasis on MAPIE. Chapter 3 outlines the methodology and data used in this study, including the characteristics of the credit default dataset and the modelling approaches employed. Chapter 4 presents the results and discusses the findings. Finally, Chapter 5 concludes the thesis, highlighting key contributions and identifying areas for future research.

2 LITERATURE REVIEW

Credit scoring plays a crucial role in financial decision-making. Traditional methods, particularly Logistic Regression, have long been the standard for credit risk evaluation due to their interpretability and regulatory acceptance (Zedda 2024). However, as financial datasets grow larger and more complex, traditional models struggle with capturing non-linear relationships and fail to leverage vast amounts of heterogeneous data (Lessmann et al. 2015). Consequently, machine learning models, particularly ensemble and boosting methods, have emerged as strong alternatives for improving predictive performance.

ML techniques have shown significant promise in improving the accuracy of predictions in credit scoring by leveraging their ability to model complex, non-linear relationships. Particularly, models like XGBoost and LightGBM outperform traditional models in predicting mortgage defaults by handling non-linearity, incorporating feature interactions, and adjusting for imbalance using boosting techniques (Zedda 2024). However, their widespread adoption has been hindered by challenges in interpretability, as they are 'black-box' models, raising concerns on whether their recommendations can be justified to managers, auditors and supervisors, or even be interpreted (Doumpos & Zopounidis 2019). Indeed, a big influence in why Logistic Regression is still widely used in credit scoring, is thanks to its interpretability, allowing for more transparent decisions, as it offers well-defined coefficients that help stakeholders understand the impact of each feature on the likelihood of default (Zedda 2024). Some alternatives to Logistic Regression are described below.

Random Forest, an ensemble of decision trees, has gained popularity due to its robustness and stability in predictive tasks. Unlike single decision trees, it is less prone to overfitting, making it well-suited for high-dimensional financial data (Li & Wu 2024).

XGBoost, a gradient boosting algorithm, has emerged as a dominant model in credit risk assessment. Unlike Random Forest, which averages multiple trees, XGBoost builds decision trees sequentially, correcting errors iteratively (Chen & Guestrin 2016). This method enables it to achieve high predictive accuracy, particularly in imbalanced datasets (Zedda 2024). Additionally, XGBoost allows the use of a parameter (*scale_pos_weight*) which assigns different weights to each class, so as to handle class imbalance, making it a preferred choice for mortgage default prediction.

LightGBM (Ke et al. 2017), another gradient boosting model, has been recognised for its efficiency and scalability, particularly in large datasets. Research suggests that LightGBM trains faster than XGBoost while maintaining comparable predictive accuracy, making it ideal for real-time financial applications (Zedda 2024). Moreover, LightGBM may employ the same parameter used by XGBoost, in order to adjust to imbalanced data.

When evaluating these models, standard metrics like accuracy are misleading due to the imbalance in the data, demanding a focus on metrics such as recall, F1-score, Receiver Operating Characteristic (ROC)-Area Under the Curve (AUC), and Precision-Recall (PR)-AUC (Lessmann et al. 2015).

Another way to get interpretability on the models, without using Logistic Regression, is by implementing feature importance metrics, like SHapley Additive exPlanations (SHAP), which provides insights into the contribution of individual variables to predictions (Mosca et al. 2022).

Studies also suggest that the integration of alternative data sources, such as transaction histories, social media activity, and psychometric information, has expanded the horizons of credit scoring (Yan 2024) (Djeundje et al. 2021).

Mortgage default prediction, a subset of credit scoring, poses additional challenges due to high class imbalance, where defaults make up a small fraction of observations. This imbalance can lead models to favour the majority class (non-default), reducing the ability to detect actual defaulters (Chen et al. 2024). Techniques such as oversampling, undersampling, and algorithmic modifications have been proposed, but their effectiveness varies across different contexts (Mohammed et al. 2020)(Verbeke et al. 2012).

While machine learning models offer high predictive power, they often lack calibration, meaning that their probability estimates do not reflect actual likelihoods; this leads to unreliable probability estimates, which are critical in financial decision-making (Gupta et al. 2022). A model is well-calibrated if, for example, a probability of 50% means that the event occurs 50% of the time (Dawid 1982). This is particularly important in financial decision-making, where well-calibrated probabilities are crucial (Silva Filho et al. 2023).

One calibration method that addresses this issue is Platt Scaling (also called Sigmoid). It applies logistic regression to model outputs, adjusting them to better reflect probabilities (Silva Filho et al. 2023). However, it is controversially discussed in the classifier calibration literature (Böken 2021). A more recent method, which tends to provide better results, is Isotonic Regression (Jiang et al. 2011). It maps model predictions to probabilities, ensuring monotonicity and improved calibration (Silva Filho et al. 2023).

To evaluate the effectiveness of calibration, researchers rely on Brier scores, which measure the mean squared difference between predicted probabilities and actual outcomes. Lower Brier scores indicate better-calibrated models. The Brier Score is an effective metric for evaluating calibration, offering a more nuanced view than traditional accuracy-based metrics (Gneiting & Raftery 2007) (Rufibach 2010).

Besides having calibrated predictions, understanding how uncertain these predictions are is essential (Jurado et al. 2015). Conformal prediction can be useful in such situations. The core idea behind it is that, instead of giving a single prediction for each row of data, a range of predictions is offered with a qualified guarantee that the true value will be within a specified range or set of options. For example, a 90% requirement guarantees that 90% of the true values will fall within the range or set given by the conformal prediction (Shafer & Vovk 2008).

MAPIE is an open-source library designed for distribution-free uncertainty quantification, including classification tasks, in a way that is model-agnostic, meaning it can work with any predictive model. It leverages conformal prediction methods to provide mathematically guaranteed prediction sets for multi-class classification problems. The MAPIE library allows users to specify the number of splits for calibration, and offers both split- and cross-conformal prediction approaches, ensuring flexibility and efficiency in generating reliable prediction sets. MAPIE's versatility in working with any machine learning model makes it particularly valuable for addressing the challenges of categorical variable prediction in imbalanced datasets (Cordier et al. 2023).

3 METHODOLOGY

Building upon the findings from the literature, this section outlines the methodology followed, using various machine learning models to predict mortgage default, with a special attention to the fact that our data is imbalanced. We will go over the topics of exploratory data analysis, data preprocessing, feature selection, model selection, model training, calibration, and uncertainty quantification using MAPIE. By following this methodology, the aim lies in producing a robust, well-calibrated model that banks can trust for lending decisions.

3.1 Dataset Description

The dataset used in this study consists of anonymized mortgage default information from a Portuguese commercial bank.

The data consists of over 75 thousand observations over the period of 2001-2008, for which the target variable is a binary indicator of credit default, where "1" represents a default and "0" represents no default. The dataset contains 39 predictor variables, 15 categorical and 24 numerical.

All the variables present in the dataset were renamed for the purpose of this analysis, as their original names were in Portuguese.

Due to the inherent nature of the problem, the dataset is highly imbalanced, having only 5.3% of its instances being "1" (cases of default). This imbalance required the use of appropriate techniques, which will be explained and discussed in further detail below.

3.2 Exploratory Data Analysis

It is essential to perform Exploratory Data Analysis before making decisions regarding our data, such as feature selection. Several statistics and visualisations for each variable in the dataset were generated. These provided a better understanding of each variable, how skewed it was, and if there was any concerning aspect about it (such as having many zero values). One of the key insights that came from this exploration was the fact that most money-related variables in the dataset were quite positively skewed.

In addition, bivariate visualisations were plotted, in order to comprehend the relationship between the target variable and the predictors.

3.3 Data Preprocessing

After performing Exploratory Data Analysis (EDA), the data preprocessing was conducted, in order to ensure that the dataset was clean, properly formatted, and ready for effective modelling and analysis.

The first step taken was to drop columns that did not add information in the presence of others. For example, the date of birth was dropped, since there was a predictor with the age at the time of scoring. The same rationale was followed to drop the employment start date, as we already had employment tenure.

The first non-payment date column was also dropped, since that only applies to cases where default is positive, so it would not make sense for an analysis in which we are trying to predict default.

Due to the fact that models are unable to process dates as features, the years and the months of the scoring and the contract dates were extracted, therefore creating four new features. In addition, a variable containing the number of days between performing the scoring and signing the contract was created, since there could be a pattern of default happening when there is a long period of time between these two events. Afterwards, these two date columns were dropped.

Another crucial step was identifying and dealing with missing values. Only one numerical column had such values, employment tenure. This column was significantly positively skewed, and studies advise against using mean imputation on skewed distributions, as that would bias the imputated values towards the tail of the distribution (M. Alwateer 2024). Therefore, the median was used, due to its robustness to outliers and skewness, and we still preferred a simpler approach, rather than a more complex one, such as K-Nearest Neighbors (KNN).

For the categorical variables with missing values, a value of *'Unknown'* was used for the property regime and the district, while a value of *'None'* was the choice taken for the type of employment contract and the professional group code, as the absence of these could mean that the person is either unemployed or retired.

In terms of handling numerical outliers, which were exclusively present in features regarding finances (such as monthly income and contracted amount), the main goal was to prevent the outliers from harming the models, while also maintaining their information, which could be useful. Therefore, the method used was Winsorization. It is a statistical technique designed to mitigate the influence of outliers, by capping extreme data points to a specified percentile, thereby reducing distortion. By replacing the highest values with less extreme ones (for a positively skewed variable), Winsorization preserves the dataset's overall structure, making it a better option than simply removing outliers (Chambers et al. 2000).

Additionally, categorical encoding was performed, using one-hot encoding, since all the categorical data was nominal. After this, our dataset had over 110 columns, meaning that performing feature selection would be essential.

Lastly, the dataset was randomly split into a 70% training set (53,014 samples) and a 30% test set (22,721 samples). It is important to perform this split before feature selection so as not to violate the principle of keeping test data unseen until the final evaluation. It must be noted that, to address the class imbalance in the data, the split was performed using stratified sampling, ensuring that each set maintained a representative distribution of the default variable (Sadaiyandi et al. 2023).

3.4 Metrics

Before diving into the Feature Selection methodology, it is important to explain the metrics utilised to search for the best model, and to evaluate our findings.

When it comes to classification problems, there are four evaluation metrics commonly used - **accuracy**, **recall**, **precision** and **F1-score**. These are called **threshold-based metrics**, as they require selecting a specific decision threshold (a number from 0 to 1) to classify predictions as positive or negative. Their formulas are the following:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Where:

- TP (True Positives): Correctly predicted positive cases.
- TN (True Negatives): Correctly predicted negative cases.
- FP (False Positives): Incorrectly predicted positive cases.
- FN (False Negatives): Incorrectly predicted negative cases.

In many classification problems, accuracy is used as the main metric, as it represents the percentage of correctly predicted instances. However, given the imbalanced nature of the data, standard accuracy was deemed insufficient as a performance metric - in our dataset, where the minority class consists of only 5.3%, a model predicting all instances as negative (majority class) would have an accuracy of over 94%, which in another context could be a great value! Therefore, there was a need to focus on other evaluation metrics.

Among the evaluated metrics, recall is particularly critical, since failing to identify a default can have a bigger financial impact than incorrectly flagging someone as a defaulter (i.e. false negatives are more harmful than false positives). However, recall should not be maximised on its own, without any other constraints or metrics taken into account, as that could make the model classify every case as positive, attaining perfect recall, while being a terrible model. Hence, the F1-score was utilised quite frequently in this analysis, as it takes recall into consideration, but also precision, making sure that there is an equilibrium.

In addition, two valuable metrics can be used to evaluate the performance of binary classifiers: **ROC-AUC** and **PR-AUC**. These are called **ranking-based metrics**, as they measure how well the model ranks positive samples higher than negative ones. They evaluate the model across all possible classification thresholds instead of a single one, in comparison to threshold-based metrics, making them more informative when dealing with imbalanced datasets, as they do not depend on a fixed threshold.

ROC-AUC quantifies how well the model distinguishes between the positive and negative classes across various thresholds. A higher AUC value (closer to 1) indicates better model performance, while a value of 0.5 suggests random guessing. The ROC-AUC is computed as the integral of the ROC curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR), as seen in 1.

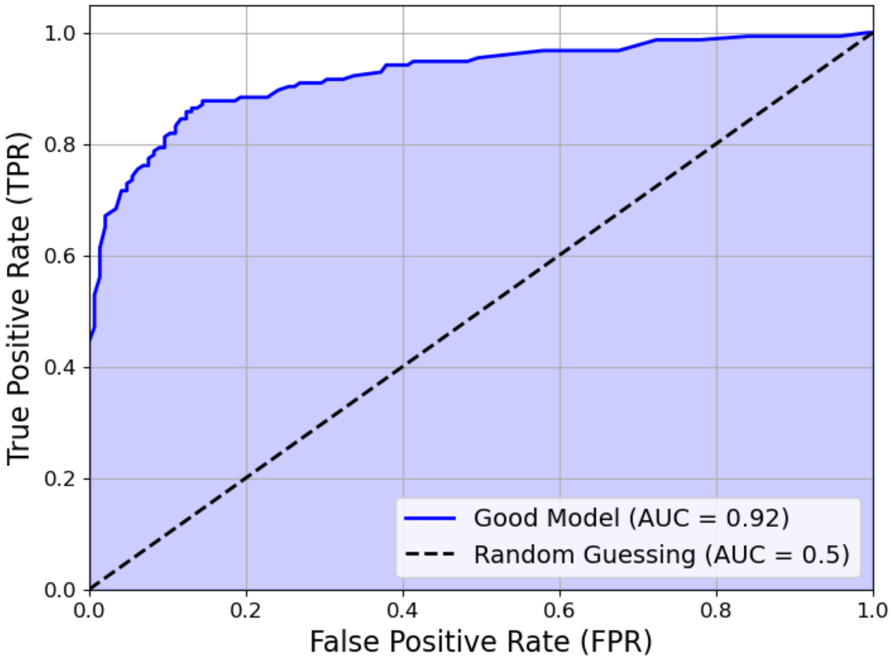

FIGURE 1: ROC-AUC example

PR-AUC measures a classification model's ability to distinguish between positive and negative classes by evaluating the trade-off between precision and recall across different threshold values, with a higher AUC indicating better performance, especially in imbalanced datasets. The baseline for this metric is the fraction of instances belonging to the minority class. The PR-AUC is computed as the integral of the Precision-Recall curve, as seen in 2.

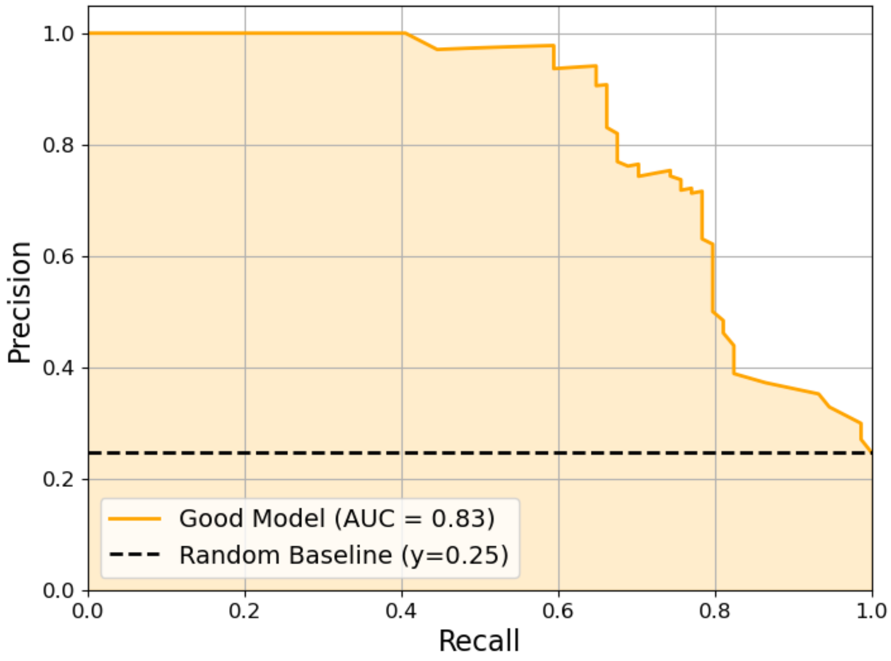

FIGURE 2: PR-AUC example

3.5 Feature Selection

Feature selection is an essential step towards improving model performance and interpretability, and reducing overfitting, particularly in a situation where we have over 100 variables. The selection process, conducted only on training data for the reasons mentioned in 3.3, aimed to eliminate redundant and less predictive features, while retaining those with the highest contribution to model performance.

3.5.1 Correlation and Variance Inflation Factor

A heatmap was plotted so as to identify variables with high correlation, and three pairs were identified: scoring and contract date (very high), monthly and annual income (very high), fixed deposits and average balance (moderate). For the pairs with very high correlation (> 0.95), only one of the variables was kept - monthly income and scoring date, as they were easier to relate with other variables (such as monthly expenses, and age at time of scoring). However, both variables from the pair with moderate correlation were kept, after calculating their Variance Inflation Factor (VIF).

The Variance Inflation Factor (VIF) for a predictor variable X_i is given by:

$$VIF_i = \frac{1}{1 - R_i^2} \tag{5}$$

where R_i^2 is the coefficient of determination obtained by regressing X_i on all other independent variables.

While correlation only shows the pairwise relationship between two variables at a time, VIF measures multicollinearity among multiple independent variables in a regression model. This formula was calculated for those two variables, and since their values were low (< 3.0) they were both kept. The VIF was also calculated for all other variables, to assure no features with high collinearity would be fed to the models, using a threshold of 5 to decide whether or not to keep the variable. After removing the features previously mentioned, no predictor had VIF > 5, hence none were removed.

After the above steps, the rest of the feature selection was performed by wielding 3 powerful tools, employing a hybrid feature selection approach:

- **SHAP Values:** Measure the contribution of each feature to the model's predictions.
- **Model-based Importance Scores from XGBoost:** Evaluate how often a feature is used in tree splits.
- **Recursive Feature Elimination (RFE):** Iteratively remove the least informative features using Logistic Regression.

3.5.2 SHAP Values

Studies have found that SHAP improves feature selection by identifying non-linear feature relationships in mortgage default risk (Li & Wu 2024). In addition, it is also great in terms of explainability, which is essential for transparent credit decision-making (Hjelkrem & Lange 2023). Hence, it was the first technique used for feature selection, and the one responsible for the greatest amount of selected features. Figure 3 depicts the SHAP Summary Plot, a great visualisation to understand the impact of each feature on model output.

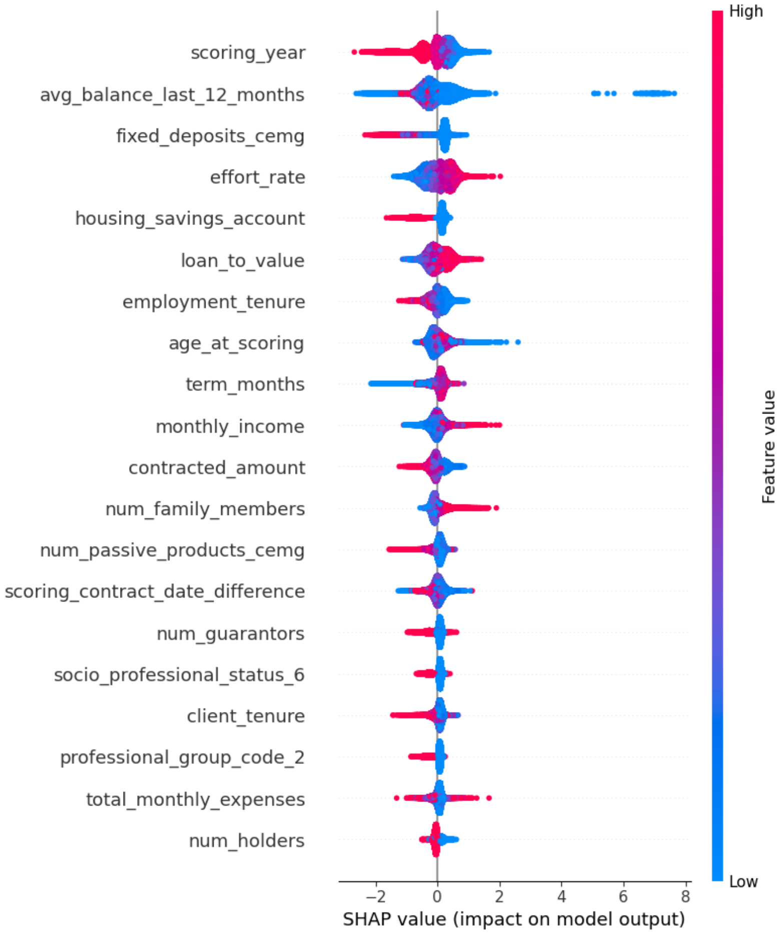

FIGURE 3: SHAP Summary Plot

Some approaches suggest taking, for example, the top 10 features in terms of SHAP value. However, deciding on a threshold seemed like a more correct approach. By setting our threshold at 0.15, features such as contracted amount and number of family members were kept which, according to domain knowledge, make sense to keep. Hence, 12 features were kept from this analysis.

3.5.3 XGBoost Feature Importance

An XGBoost classifier was trained on the dataset, using 5-fold Cross-Validation (CV). After training the model, the feature importance was extracted using the built-in *feature_importances_* attribute, which ranks the features based on their contribution to the model's predictive power. Figure 4 shows the obtained results.

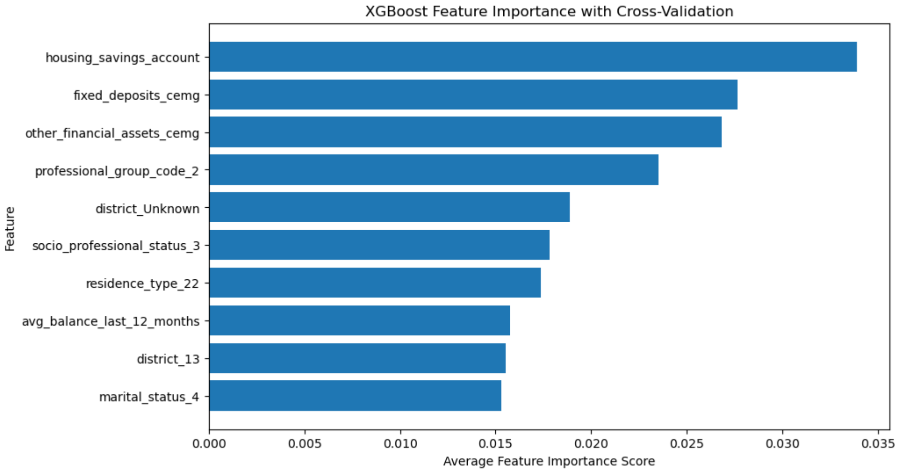

FIGURE 4: XGBoost Feature Importance

This model seems to give more importance to categorical variables contrasting with the results obtained from SHAP, as shown in table I, where the outcomes from SHAP and XGBoost are compared.

TABLE I: Comparison of Feature Importance: SHAP vs. XGBoost

Feature	**Mean SHAP Value**	**XGBoost Importance**
scoring_year	0.391	-
avg_balance_last_12_months	0.386	0.016
fixed_deposits_cemg	0.367	0.028
effort_rate	0.343	-
housing_savings_account	0.256	0.034
loan_to_value	0.248	-
employment_tenure	0.189	-
age_at_scoring	0.184	-
term_months	0.176	-
monthly_income	0.167	-
other_financial_assets_cemg	-	0.027
professional_group_code_2	-	0.024
district_Unknown	-	0.019
socio_professional_status_3	-	0.018
residence_type_22	-	0.017

Note that the values are omitted when they are low (for SHAP < 0.15 and for XGBoost < 0.015). Interestingly, not many variables were considered important by both approaches, only the average balance, fixed deposits (both numerical) and housing savings account (binary indicator).

Besides the already selected features from SHAP, only *other_financial_assets_cemg* and *professional_group_code_2* were kept from XGBoost feature importance results, as they are the only ones with values above 0.02.

3.5.4 Recursive Feature Elimination (RFE)

The third method was RFE, done through a simple logistic regression model. RFE works by recursively removing the least important features based on the model's performance until the optimal subset of features of size n is identified. The results performed using n = 10 are presented in table II.

TABLE II: Presence of RFE Selected Features in Baseline Model

RFE Selected Feature	**Present in Baseline Model**
housing_savings_account	✓
employment_contract_type_03	×
employment_contract_type_None	×
socio_professional_status_10	×
socio_professional_status_12	×
professional_group_code_2	✓
district_12	×
district_44	×
district_45	×
district_46	×

3.5.5 Final Feature Set

The SHAP + XGBoost feature selection set served as the baseline model. RFE-selected features were then tested incrementally to determine their impact when added to the baseline model. The metric used to determine the performance of each Feature Set was the ROC-AUC, which is used to evaluate the performance of binary classification models by measuring their ability to distinguish between positive and negative classes. The ROC curve can be defined as a plot of True Positive Rate (Recall) vs. False Positive Rate at various classification thresholds, while AUC consists of a single value, representing the total area under the ROC curve. The higher the AUC, the better the model is at distinguishing between the two classes.

To identify the optimal combination of RFE features to be appended, an exhaustive search was conducted, using the 8 features not previously selected, and evaluating all possible subsets when added to the baseline model. After computing the ROC-AUC score for the baseline model, all possible feature combinations were systematically tested, combining individual ones as well as groups of two, three, four, and so on, with the ones from SHAP + XGBoost, looking for the combination yielding the best ROC-AUC value.

After testing all RFE feature combinations, the optimal subset that yielded the highest ROC-AUC consisted of only two additional features: *employment_contract_type_None* and *district_46*. The various ROC-AUC values are compared in table III.

Table III: ROC-AUC Comparison Across Feature Selection Approaches

Feature Selection Strategy	**ROC-AUC**
SHAP + XGBoost (Baseline)	0.7904
RFE	0.6327
SHAP + XGBoost + RFE	0.7919
SHAP + XGBoost + Optimal RFE Feature Combination	**0.7942**

Final Selected Features:

- avg_balance_last_12_months
- fixed_deposits_cemg
- effort_rate
- housing_savings_account
- loan_to_value
- employment_tenure
- age_at_scoring
- term_months
- monthly_income
- num_family_members
- contracted_amount
- scoring_year
- professional_group_code_2
- other_financial_assets_cemg
- employment_contract_type_None
- district_46

The final predictive model, containing 16 features, managed to balance complexity and interpretability while achieving the highest ROC-AUC. This study demonstrates that a hybrid feature selection approach, combining SHAP, XGBoost, and targeted RFE selection, can improve model performance while maintaining interpretability.

3.6 Model Selection

This study is focused on a binary classification problem, in the subject of financial risk assessment, with an imbalanced dataset. Therefore, those factors were taken into account when choosing which models to apply to the data.

Logistic Regression was the first model selected as it is a widely used statistical method in credit scoring due to its simplicity and interpretability (Zedda 2024). Studies have demonstrated its utility in predicting loan defaults, making it a reliable baseline model. Logistic Regression is computed through the following formula:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)}} \tag{6}$$

Where:

- $P(Y = 1 \mid X)$ is the probability that the outcome Y is 1 (the positive class);
- β_0 is the intercept term;
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients for the input features $X_1, X_2, \ldots, X_n$;
- X represents the input features.

This formula provides direct probability estimates, which are useful for understanding the influence of each variable.

Then, because of the imbalanced target variable, tree-based and boosting models were considered due to their ability to handle skewed distributions. Indeed, these methods have proved highly effective in modelling non-linear data when working with financial data (Xuan et al. 2018) (Li et al. 2020).

Random Forest (Breiman 2001) was chosen because of its ability to capture complex interactions between features, as well as being well-suited for handling imbalanced data, since it possesses a built-in parameter dedicated to that. Random Forest constructs multiple decision trees to improve predictive performance and control overfitting. Its ability to handle high-dimensional data and capture complex interactions makes it suitable for financial datasets (Li & Wu 2024), such as this one. Additionally, research indicates that Random Forest models outperform single decision trees (Breiman 2001) in loan default predictions (Li & Wu 2024). The formula for Random Forest is the following:

$$\hat{Y} = \frac{1}{N} \sum_{i=1}^{N} T_i(X) \tag{7}$$

- $\hat{Y}$ is the predicted output;
- N is the number of decision trees in the forest;
- $T_i(X)$ is the output of the i-th decision tree for input X.

For classification, such as this case, the final prediction consists of the majority vote from all decision trees.

Gradient boosting models, XGBoost and LightGBM, were included due to their strong predictive performance, particularly when it comes to credit scoring. They are also able to handle class imbalance through built-in weighting mechanisms, and able to handle large-scale datasets better than other models.

XGBoost can achieve strong performance in credit scoring applications, particularly when it comes to distinguishing between good borrowers and potential defaulters, often surpassing traditional methods (Zedda 2024). On the other hand, LightGBM is noted for its speed and scalability when dealing with large-scale financial data (Li & Wu 2024).

Studies have shown that models built using XGBoost and LightGBM achieve higher predictive power compared to logistic regression and decision tree models (Zedda 2024).

This approach leverages each model's respective strength - the interpretability of Logistic Regression, the robustness of Random Forest, XGBoost's sensitivity, and the scalability of LightGBM - providing a comprehensive approach to mortgage default prediction.

3.7 Model Training

As previously mentioned in section 3.3, the data has been randomly split into a 70% training set (53,014 samples) and a 30% test set (22,721 samples). Before training the model, the test set was split in half (11,361 samples), in order to create a validation set (11,360 samples), meaning that each set used 15% of the samples. As it was explained before, both these splits used stratified sampling, to ensure that the original distribution of the target variable was maintained on all sets (Sadaiyandi et al. 2023). Each set has the following purpose:

- **Training Set:** Model training and hyperparameter tuning (using grid search with cross-validation);
- **Validation Set:** Threshold optimisation;
- **Test Set:** Final performance evaluation.

A separate DataFrame was created in order to apply feature scaling to it. This dataset was used exclusively by the Logistic Regression model, due to its sensitivity to differences in feature scales, as it relies on linear coefficients (6). In contrast, tree-based models like Random Forest, XGBoost, and LightGBM do not require scaling, since they use decision rules based on feature splits. Aside from this difference, the rest of the model training process was the same for the four models.

In terms of model-specific parameters (asides from hyperparameters), it is worth mentioning that Logistic Regression used the 'Library for Large Linear Classification' solver, as it is well-suited for binary classification tasks and for medium-sized datasets. XGBoost, on the other hand, was performed using the log-loss function, because Logarithmic Loss is the most adequate evaluation metric for binary classification.

With the goal of enhancing model performance, hyperparameter tuning was conducted using grid search with cross-validation. This process involves testing all possible combinations of specified hyperparameters, to identify the optimal one, yielding the best performance. Because of the imbalance of the data, accuracy is not an adequate metric to evaluate models, therefore, the F1-score was used to search for the best hyperparameter combinations. The training set was used, since cross-validation was applied, meaning that it automatically splits the training data into n folds, using $n - 1$ folds for training and 1 fold for validation. The hyperparameters tuned for each model were the following:

- **Logistic Regression:** *C* (controls regularisation; larger *C* $\rightarrow$ less regularisation).
- **Random Forest:** number of trees, maximum depth, minimum samples to split, minimum samples per leaf.
- **XGBoost:** number of trees, maximum depth, learning rate, subsample ratio, column sampling by tree.
- **LightGBM:** number of trees, maximum depth, learning rate, number of leaves.

With the models trained using the best hyperparameters, it was already possible to move on to the evaluation phase. However, this would mean using the default threshold of 0.5, which is often suboptimal, particularly when dealing with imbalanced data. Hence, using the validation set, threshold optimisation was conducted. Two strategies were applied:

- selecting the threshold that maximized the F1-score;
- optimising F1-score while ensuring a minimum recall of 0.8 for the positive class, prioritising the detection of default cases.

The PR-AUC was calculated for each model and plotted for better understanding of the trade-off between precision and recall across different thresholds. After analysing this, it was decided that the second option (guaranteeing a recall of at least 0.8) was more appropriate for our goal of avoiding false negatives, and so it is the approach followed before computing metrics on the test set. The optimised thresholds were plotted on the corresponding curve, serving as an aid for understanding the logic behind their selection.

After all the steps explained above, the confusion matrix and the classification report were generated, so as to determine how well the models performed.

3.7.1 Handling Data Imbalance

When dealing with imbalanced data, models tend to favor the prediction the majority class (*default* = 0), if no appropriate strategies to deal with imbalance are followed. Therefore, two separate approaches were implemented.

The first one consisted of using the built-in parameter Class Weighting, for both Logistic Regression and Random Forest, set to *balanced*, while for XGBoost and LightGBM Class Weighting for Positive Class was used, which can be obtained from the formula $\frac{\text{Number of negative samples}}{\text{Number of positive samples}}$. Both of these are meant to make the model "pay more attention" to the minority class, penalising more harshly its misclassification (i.e. predicting 0 when the true value is 1 is more penalised than predicting 1 when the true value is 0).

The second strategy consisted of performing oversampling of the minority class, as the literature indicates that oversampling methods generally outperform undersampling methods, particularly, the most significant performance gains are brought by using the Synthetic Minority Oversampling Technique (SMOTE) algorithm (Haluska et al. 2022). Hence, SMOTE was used to oversample the minority class (default cases) before model training, creating a new dataset with around 100,000 samples. However, this process might generate noisy samples, which must be cleaned for better model performance. So, instead of performing SMOTE by itself, Tomek Links were also used to obtain a cleaner sample space. Indeed, the literature supports the use of a hybrid resampling approach for extremely imbalanced data (Wongvorachan et al. 2023).

The goal of the Tomek Links, an under-sampling technique, is to eliminate some of the default cases that are near the edges of, or are surrounded by, the set of non-default cases, in order to define a clearer boundary between the majority and minority classes.

Both strategies are compared in section 4.

3.8 Calibration

When it comes to real-world credit scoring, decisions are made based on predicted probabilities of default, not just binary classifications. A model may correctly rank a customer as high risk, but still assign an incorrect probability value to them. Calibration ensures that a predicted probability of 0.5, for example, would truly correspond to a 50% chance of default, hence guaranteeing probabilities that are reliable and interpretable for decision-making.

In addition, since the target variable is imbalanced, models tend to be overconfident in predicting the majority class. Calibration adjusts these probability estimates, making them more aligned with actual default rates. This is extremely important because, in credit risk assessment, lenders set thresholds for loan approvals based on probability estimates. Poor calibration on imbalanced data usually leads to underestimating high-risk borrowers, resulting in financial losses.

Moreover, MAPIE's implementation, detailed in section 3.10, relies on probability distributions to construct prediction intervals. If the probabilities are not calibrated, the prediction intervals may be too wide or too narrow, affecting coverage reliability. Proper calibration ensures accurate confidence intervals.

In terms of calibration methods, Platt Scaling and Isotonic Regression were considered. The first is known for working well when the relationship between the predicted scores and actual probabilities is sigmoidal, making it ideal for Logistic Regression models. Isotonic Regression, on the other hand, is more flexible than Platt scaling, as it does not assume a specific functional form, and works well when the mapping between raw scores and true probabilities is non-linear but monotonic, making it more adequate for Random Forest and boosting models. Despite these differences, both calibration methods were applied to all the models. The relevant plots were generated as the effectiveness of calibration is easier to understand visually.

Additionally, the Brier Scores were calculated for each model (for the uncalibrated models, and for each of the calibration methods). The Brier Score is used for evaluating how well a model predicts probabilities in binary classification problems, in other words, it measures the accuracy of probabilistic predictions (Bradley et al. 2008). The lower the Brier score, the better the probabilistic predictions - a perfectly calibrated model has a Brier Score of zero.

The Brier Score (BS) is calculated as:

$$BS = \frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_i - Y_i)^2 \tag{8}$$

- BS is the Brier score, measuring the accuracy of probabilistic predictions.
- N is the total number of predictions.
- $\hat{Y}_i$ is the predicted probability of the positive class for instance i.
- Y_i is the actual observed class for instance i, where $Y_i \in \{0, 1\}$.

The calibration results, plots and comparisons can be found in section 4.2

3.9 Evaluation

For a broader look at all the models, both with and without SMOTE oversampling, and for the sake of easily detecting the strengths and weaknesses of each model, a table was generated displaying many metrics of interest, namely the typical classification metrics - accuracy, precision, recall and F1-score - along with other relevant metrics - ROC-AUC, PR-AUC, the 3 Brier Scores, and the threshold used.

3.10 MAPIE

The MAPIE framework was applied in this study to estimate prediction intervals and assess uncertainty in the predictions. MAPIE does this by providing confidence intervals for predicted class labels, ensuring calibrated predictions. In a binary classification problem, such as this one, there are only 3 possible prediction intervals: $[0, 0]$; $[0, 1]$; $[1, 1]$

Essentially, for a confidence interval of 90$\%$ (for example), the model will output [0, 0] if it is at least 90$\%$ sure that the sample belongs to the negative class. Similarly, the model will choose an interval of [1, 1] if it is at least 90$\%$ sure that it is handling a positive class. On the other hand, if MAPIE is not at least 90$\%$ sure about either class, it will then output [0, 1].

The MAPIE implementation was achieved through a custom function, which was called for each 'best model', i.e. each model after hyperparameter tuning, along with the training data for both the predictors and the target variables, and the test data for the predictors. Besides those parameters, it also takes:

- α (alpha): the significance level for the prediction intervals (with a default value of 0.1), ensuring a coverage probability of 1 - α.

- method: the approach used to estimate conformal prediction intervals; only two MAPIE methods are supported for binary classification tasks:
  - "score": the standard conformal prediction method that estimates prediction sets based on score functions;
  - "lac" (Least Ambiguous set-valued Classifier (LAC)): a method designed to improve the calibration of probability-based predictions.
- cv (Cross-Validation strategy): determines how MAPIE estimates prediction intervals; the following options were explored, for the sake of comparing results:
  - cv=None: no cross-validation, relying on a single training pass;
  - cv=5: a five-fold cross-validation strategy to increase robustness in uncertainty estimation;
  - cv="prefit": MAPIE was applied to models that were pre-trained, meaning no additional training was performed; when using this option, the method called is indifferent, the result will be the same.

This function returns the predictions, the intervals, and the upper and lower bounds (calculated from the intervals) for each test sample.

Each of the five possible combinations (the score and LAC methods combined with no cross validation and 5-fold CV, in addition to the "prefit" option) was tested across all models, including those that had been previously calibrated, so as to assess the impact of probability calibration on conformal prediction intervals. This approach ensured a comprehensive evaluation of different prediction interval estimation techniques, allowing for a comparative analysis of the effect of cross-validation and calibration in improving uncertainty quantification.

3.10.1 Evaluation for MAPIE

In similar fashion, a function with the purpose of calculating metrics on the MAPIE output was created. It receives the results from calling the previously described function, as well as the true values of the target variable on the test set, and returns the following metrics:

- Coverage Rate:
  - This measures the proportion of true outcomes that fall within the prediction interval.
  - A higher coverage rate means the model's prediction interval successfully captures the true value more often.
  - Formula:

    $$\text{Coverage Rate} = \frac{1}{n}\sum_{i=1}^{n} 1\left(Y_i \in [Y_{\text{lower},i}, Y_{\text{upper},i}]\right) \tag{9}$$

    Where:

    * n is the total number of samples;
    * Y_i is the true label for sample i;
    * $Y_{\text{lower},i}$ and $Y_{\text{upper},i}$ are the lower and upper bounds of the prediction interval for sample i;
    * $1\ (\cdot)$ is an indicator function that returns 1 if the true label Y_i is within the prediction interval, and 0 otherwise.

- Miscoverage Rate:
  - This measures the proportion of true outcomes that fall outside the prediction interval.
  - Lower miscoverage rates are preferable.
  - Formula:

    $$1 - CoverageRate \tag{10}$$

- Average Interval Width (AIW):
  - This represents the average width of the prediction interval.
  - Smaller interval widths are generally preferred, as they indicate more precise predictions.

  - In binary classification, AIW represents the percentage of prediction intervals that are [0, 1].
  - This metric should be analysed along with the coverage rate to balance uncertainty quantification with informativeness, ensuring that prediction intervals are neither too wide (overly conservative) nor too narrow (overconfident and unreliable).
  - Formula:

$$\text{AIW} = \frac{1}{N}\sum_{i=1}^{N}(U_i - L_i) \tag{11}$$

    Where:

    * N is the total number of predictions;
    * U_i is the upper bound of the prediction interval for instance i;
    * L_i is the lower bound of the prediction interval for instance i.

- Asymmetry:
  - This metric quantifies the asymmetry of the prediction intervals, indicating whether the intervals are balanced around the predicted values or skewed towards one side.
  - A positive asymmetry value means the upper bound is wider than the lower bound, whereas a negative asymmetry value means the opposite.
  - A perfectly symmetric prediction interval would have an asymmetry value close to zero.
  - Formula:

$$\text{Asymmetry} = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_{\text{upper},i} - \hat{Y}_i}{\hat{Y}_i - Y_{\text{lower},i}} - 1 \tag{12}$$

    Where:

    * n is the total number of samples;
    * $\hat{Y}_i$ is the predicted value for sample i;
    * $Y_{\text{lower},i}$ and $Y_{\text{upper},i}$ are the lower and upper bounds of the prediction interval for sample i.

There was a need for a metric that would take into account how many predictions are correctly covered, and how wide the intervals to get to those predictions were - correct predictions that contain both classes do not give us much information, aside from the uncertainty. Hence, Exact Match Rate was added.

- Exact Match Rate:
  - This metric measures the proportion of predictions where the true outcome **exactly** matches both the lower and upper bounds of the prediction interval, a metric that only makes sense in classification problems.
  - A higher Exact Match Rate indicates that the model is highly confident in its predictions, producing narrower intervals that precisely capture the true value.
  - In binary classification problems, this can simply be calculated by subtracting the Interval Width from the Coverage Rate.
  - Formula:

$$\text{Exact Match Rate} = \frac{1}{n}\sum_{i=1}^{n} 1\left(Y_i = Y_{\text{lower},i} = Y_{\text{upper},i}\right) \tag{13}$$

    Where:

    * n is the total number of samples;
    * Y_i is the true label for sample i;
    * $Y_{\text{lower},i}$ and $Y_{\text{upper},i}$ are the lower and upper bounds of the prediction interval for sample i;
    * $1\ (\cdot)$ is an indicator function that returns 1 if the true label Y_i is equal to both bounds, and 0 otherwise.

4 RESULTS

In this section, the results of all the steps performed in the methodology section are presented, so as to evaluate the different machine learning models and techniques applied to mortgage default prediction.

Firstly, the models' performances are compared in terms of threshold-based metrics, as well as ranking-based metrics. Then, both calibration methods are compared. Lastly, the results of uncertainty quantification using MAPIE are analysed.

4.1 Model Comparison

A model comparison is shown in table IV. Note that when a model is labelled *"SMOTE"*, not only SMOTE was applied to that table, but also Tomek Links, as previously explained in section 3.7.1.

TABLE IV: Model Evaluation Results

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
Logistic Regression	0.506	0.082	0.815	0.150
Logistic Regression (SMOTE)	0.501	0.082	0.825	0.149
Random Forest	0.451	0.082	0.916	0.151
Random Forest (SMOTE)	0.608	0.103	0.831	0.184
XGBoost	0.616	0.103	0.806	0.183
XGBoost (SMOTE)	0.607	0.099	0.786	0.175
LightGBM	0.601	0.098	0.788	0.174
LightGBM (SMOTE)	0.597	0.097	0.788	0.172

Model	**ROC-AUC**	**PR-AUC**	**Threshold Used**
Logistic Regression	0.724	0.156	0.043
Logistic Regression (SMOTE)	0.723	0.146	0.430
Random Forest	0.767	0.180	0.297
Random Forest (SMOTE)	0.788	0.172	0.123
XGBoost	0.792	0.304	0.053
XGBoost (SMOTE)	0.769	0.210	0.002
LightGBM	0.778	0.231	0.129
LightGBM (SMOTE)	0.768	0.239	0.006

The statistical differences between models are not very significant. Nevertheless, it is worth comparing them.

Overall, the best model seems to be XGBoost, as it achieves the highest ROC-AUC (0.792) and PR-AUC (0.304), and the second highest F1-Score (0.183), indicating that it is the best model for ranking defaulters correctly. The fact that it used the Class Weighting for Positive Class parameter to handle imbalanced data could mean that this consists of a powerful solution in the challenge of data imbalance.

It can also be concluded that Random Forest benefits from the use of SMOTE. While Random Forest without SMOTE has high recall (0.916), Random Forest with SMOTE performs better in all the other threshold-based metrics. Applying SMOTE significantly improves the balance between precision and recall, therefore getting a higher F1-Score.

LightGBM had a decent performance, getting the second and third best glspr-AUC scores, and not having any worst scores in any metric. In terms of ranking, both LightGBM models (with and without SMOTE) would be in third place, behind XGBoost and Random Forest (with SMOTE). This is understandable, as LightGBM is a less computationally expensive model.

Logistic Regression seems to be the weakest model. It struggles to capture complex patterns in the data, with the lowest ROC-AUC and PR-AUC scores. While it maintains high recall, its precision is too low for reliable predictions.

These findings suggest that boosting methods (XGBoost and LightGBM) naturally handle imbalanced data well - in a situation with limited computer power or time, or when handling enormous amounts of data, LightGBM would be advised, whereas in a situation that allows the use of XGBoost, then it would be preferable. In addition, these results indicate that, if using a Random Forest model on imbalanced data, it is advised to use oversampling techniques.

4.2 Calibration

The Brier Scores for uncalibrated models, as well as the Brier Scores for the models after applying Platt Scaling and Isotonic Regression were calculated, as explained in section 3.8. The obtained values can be seen in table V.

TABLE V: Comparison of Brier Scores for different Calibration methods

Model	**Uncalibrated**	**Platt Scaling**	**Isotonic Reg.**
Logistic Regression	0.217	0.048	0.048
Logistic Regression (SMOTE)	0.215	0.048	0.217
Random Forest	0.159	0.048	0.047
Random Forest (SMOTE)	0.068	0.061	0.060
XGBoost	0.058	0.044	0.043
XGBoost (SMOTE)	0.052	0.059	0.056
LightGBM	0.088	0.046	0.045
LightGBM (SMOTE)	0.048	0.053	0.051

Several inferences can be made from the observation of the table V. Clearly, in most cases, the use of calibration improves the Brier Score significantly - recall that the Brier Score is better when it is closer to zero. This confirms that raw probability outputs from machine learning models tend to be poorly calibrated, particularly in imbalanced classification problems.

With the exception of Logistic Regression (With SMOTE), Isotonic Regression consistently slightly outperforms Platt Scaling, as it yields the lowest Brier score. This result aligns with expectations, as these models, aside from Logistic Regression, do not inherently produce sigmoidal probability distributions.

Even though Isotonic Regression tends to outperform Platt Scaling, sometimes it does not yield better results than uncalibrated models, as is the case with most of the models using SMOTE - Logistic Regression, XGBoost and LightGBM.

It can be concluded that, for Logistic Regression, it is safer to use Platt Scaling, as it performs consistently for this model. Whereas for tree-based algorithms, in which no oversampling or undersampling techniques were performed, Isotonic Regression is clearly the best option. For the rest of the models (i.e. tree based algorithms with over or undersampling methods) calibration can be a tricky endeavour, as it may not provide improvements. On the other hand, these models "redeem" themselves by having good uncalibrated Brier Scores.

4.3 MAPIE

As described in section 3.10, MAPIE was implemented for the 8 models that have been discussed so far in this thesis. It was implemented using all combinations of 3 different *cv* parameters, and 2 different MAPIE methods (asides from *cv="prefit"*, which gives the same result no matter the method). In addition, MAPIE was also implemented for the calibrated models, resulting in obtaining metrics for 120 applications of MAPIE. Given this large quantity, these results will not be presented in full, only the most noteworthy.

Because the problem in hand is a high stakes decision (failing to predict a default outcome can be very costly), the confidence interval used for all cases was of 90%.

The best coverage rate obtained was on two implementations of **Isotonic Logistic Regression**, one using *method="lac"* and *cv=None*, while the other used *method="score"* and *cv=5*. Their metrics can be seen in table VI.

TABLE VI: Models yielding Best Coverage Rate

Method	**Coverage Rate**	**Miscoverage Rate**	**AIW**	**Asymmetry**	**Exact Match Rate**
LAC	0.991	0.009	0.950	0.949	0.041
Score + CV 5	0.991	0.009	0.949	0.948	0.042

The best AIW, i.e. the tightest / smallest, was found on the model **Logistic Regression with SMOTE** calibrated using **Isotonic Regression**, and using *cv="prefit"*, as presented in table VII.

TABLE VII: Models yielding Best Average Interval Width

Method	**Coverage Rate**	**Miscoverage Rate**	**AIW**	**Asymmetry**	**Exact Match Rate**
prefit	0.330	0.670	0.193	0.723	0.138

Analysing both tables (VI and VII) leads to a clear conclusion - looking for either the best coverage rate, or the best AIW, while disregarding the other, does not lead to the best application of MAPIE. There is not much use in a models that is uncertain almost 95% of the time, nor in a model that is only correct 33% of the time.

Therefore, the model with the best exact match rate must be found, as that indicates that the model has a good balance between Coverage Rate and AIW. As previously mentioned, the oversampling and undersampling techniques (SMOTE and Tomek Links) were able to improve the **Random Forest** model significantly and, interestingly enough, that is also the model that managed to yield the best exact match rate. Similarly to the model with the best AIW, this one also used *cv="prefit"*. The metrics can be seen in table VIII.

TABLE VIII: Models yielding Best Exact Match Rate

Method	**Coverage Rate**	**Miscoverage Rate**	**AIW**	**Asymmetry**	**Exact Match Rate**
prefit	0.958	0.042	0.651	0.553	0.307

The **Random Forest model with SMOTE** manages to balance a great coverage rate (almost 96%) with an acceptable AIW (65%).

4.3.1 Effect of Calibration on MAPIE

Due to the large amount of data, it would be quite complicated to try to analyse the effect of calibration on MAPIE by looking at each case individually. Hence, the average metrics were computed for each calibration method, as well as for the models with no calibration, as presented in table IX.

TABLE IX: Comparison of Calibration Methods

Calibration	**Coverage Rate**	**Miscoverage Rate**	**AIW**	**Asymmetry**	**Exact Match Rate**
Isotonic	0.968	0.032	0.915	0.922	0.053
Sigmoid	0.984	0.016	0.934	0.927	0.050
No Calibration	0.854	0.146	0.762	0.821	0.092

The calibrated results are similar to each other, having very high coverage rates (above 96%), but also outputting huge AIW (above 91%). The models without calibration achieve more balanced metrics, covering about 85% of the data, with AIW of 76%, therefore, having a better exact match rate.

5 CONCLUSION

This study aimed to enhance mortgage default prediction by applying various machine learning techniques and addressing key challenges such as class imbalance, model calibration, and predictive uncertainty. Even though Logistic Regression remains widely used in the subject area, the findings of this research demonstrate that advanced machine learning models, particularly boosting techniques like XGBoost and LightGBM, provide superior predictive performance in distinguishing high-risk borrowers from low-risk ones.

A critical issue addressed in this thesis was the inherent class imbalance in mortgage default datasets, which can lead to biased models favouring the majority class. Two primary approaches were explored: leveraging built-in weighting mechanisms in models and using hybrid resampling techniques (SMOTE combined with Tomek Links). While the boosting models handled imbalance effectively using built-in class weighting, Random Forest showed substantial improvement when combined with the hybrid resampling techniques, emphasizing the importance of data preprocessing strategies tailored to each model type.

Beyond improving classification performance, the study also investigated probability calibration techniques. The results revealed that tree-based models benefit most from Isotonic Regression, whereas Logistic Regression performs best with Platt Scaling. Interestingly, Isotonic Regression performed better when oversampling techniques were not used.

Uncertainty quantification was another key focus of this research, with MAPIE applied to assess confidence in model predictions. The results indicate that, while MAPIE provides valuable prediction intervals, the trade-off between interval width and coverage must be carefully managed. The highest coverage rate was achieved using MAPIE with Isotonic Regression-calibrated Logistic Regression, but this came at the cost of excessively wide intervals. Another metric, named Exact Match Rate was used in order to balance both coverage rate and AIW. Using this, it was concluded that the best balance between predictive accuracy and uncertainty estimation was found with Random Forest combined with SMOTE and MAPIE, highlighting its potential for real-world application.

From a practical standpoint, these findings have some implications for financial institutions seeking to refine their credit risk assessment frameworks. The use of machine learning models, coupled with robust calibration and uncertainty estimation techniques, can enhance lenders' ability to make informed, data-driven decisions under uncertainty. However, it is essential to acknowledge the limitations of predictive modelling in mortgage default forecasting. Real-world lending decisions are influenced by external macroeconomic factors such as inflation, unemployment rates, and economic crises - factors that were not explicitly modelled in this study. Additionally, borrower behaviour can be influenced by psychological and sociopolitical dynamics that may not be fully captured in structured financial datasets.

Future research could explore the integration of macroeconomic indicators into mortgage default prediction models to account for external shocks and market conditions. Experimenting these same techniques on different mortgage default datasets, could lead to interesting results, as this dataset is missing information that could be useful, such as the borrower's credit score and the interest rate agreed to on the contract, as well as if it is fixed or variable. Additional balancing techniques could also be a good subject for future work.

Furthermore, the implementation of Venn-Abbers predictors (Vovk & Petej 2012) instead of MAPIE could also be beneficial, as they typically perform better with classification tasks.

In conclusion, while machine learning models, particularly boosting techniques and ensemble methods, demonstrate superior performance over traditional approaches, the effective handling of data imbalance, probability calibration, and uncertainty estimation is crucial for their practical applicability. By refining these aspects, financial institutions can move towards more reliable and transparent credit risk assessment models, ultimately improving the robustness of mortgage default prediction in high-stakes financial environments.

REFERENCES

Bradley, A. A., Schwartz, S. S. & Hashino, T. (2008), 'Sampling uncertainty and confidence intervals for the brier score and brier skill score', *Weather and Forecasting* **23**(5), 992–1006.

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

Böken, B. (2021), 'On the appropriateness of platt scaling in classifier calibration', *Information Systems* **95**, 101641.
URL: *https://www.sciencedirect.com/science/article/pii/S0306437920301083*

Chambers, R., Kokic, P., Smith, P. & Cruddas, M. (2000), 'Winsorization for identifying and treating outliers in business surveys', *Proceedings of the Second International Conference on Establishment Surveys* pp. 717–726.

Chen, T. & Guestrin, C. (2016), Xgboost: A scalable tree boosting system, *in* 'Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining', pp. 785–794.

Chen, Y., Calabrese, R. & Martin-Barragan, B. (2024), 'Interpretable machine learning for imbalanced credit scoring datasets', *European Journal of Operational Research* **312**(1), 357–372.
URL: *https://www.sciencedirect.com/science/article/pii/S0377221723005088*

Cordier, T., Blot, V., Lacombe, L., Morzadec, T., Capitaine, A. & Brunel, N. (2023), Flexible and systematic uncertainty estimation with conformal prediction via the mapie library, *in* 'Conformal and Probabilistic Prediction with Applications', PMLR, pp. 549–581.

Dawid, A. P. (1982), 'The well-calibrated bayesian', *Journal of the American Statistical Association* **77**(379), 605–610.

Djeundje, V. B., Crook, J., Calabrese, R. & Hamid, M. (2021), 'Enhancing credit scoring with alternative data', *Expert Systems with Applications* **163**, 113766.

Doumpos, Michalis, C. L. D. N. & Zopounidis, C. (2019), *Analytical Techniques in the Assessment of Credit Risk*, EURO Advanced Tutorials on Operational Research.

Gneiting, T. & Raftery, A. E. (2007), 'Strictly proper scoring rules, prediction, and estimation', *Journal of the American Statistical Association* **102**(477), 359–378.

Gupta, C., Podkopaev, A. & Ramdas, A. (2022), 'Distribution-free binary classification: prediction sets, confidence intervals and calibration'.
URL: *https://arxiv.org/abs/2006.10564*

Haluska, R., Brabec, J. & Komarek, T. (2022), Benchmark of data preprocessing methods for imbalanced classification, *in* '2022 IEEE International Conference on Big Data (Big Data)', IEEE, p. 2970–2979.
URL: *http://dx.doi.org/10.1109/BigData55660.2022.10021118*

He, H. & Garcia, E. A. (2009), 'Learning from imbalanced data', *IEEE Transactions on Knowledge and Data Engineering* **21**(9), 1263–1284.

Hjelkrem, L. O. & Lange, P. E. d. (2023), 'Explaining deep learning models for credit scoring with shap: A case study using open banking data', *Journal of Risk and Financial Management* **16**(4).
URL: *https://www.mdpi.com/1911-8074/16/4/221*

Jiang, X., Osl, M., Kim, J. & Ohno-Machado, L. (2011), 'Smooth isotonic regression: A new method to calibrate predictive models', *AMIA Joint Summits on Translational Science Proceedings* pp. 16–20.

Jurado, K., Ludvigson, S. C. & Ng, S. (2015), 'Measuring uncertainty', *American Economic Review* **105**(3), 1177–1216.

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**.

Lessmann, S., Baesens, B., Seow, H.-V. & Thomas, L. C. (2015), 'Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research', *European Journal of Operational Research* **247**(1), 124–136.
URL: *https://www.sciencedirect.com/science/article/pii/S0377221715004208*

Li, H., Cao, Y., Li, S., Zhao, J. & Sun, Y. (2020), 'Xgboost model and its application to personal credit evaluation', *IEEE Intelligent Systems* **35**(3), 52–61.

Li, H. & Wu, W. (2024), 'Loan default predictability with explainable machine learning', *Finance Research Letters* **60**, 104867.
URL: *https://www.sciencedirect.com/science/article/pii/S1544612323012394*

M. Alwateer, E. Atlam, M. E.-R. O. G. I. G. (2024), 'Missing data imputation: A comprehensive review', *Journal of Computer and Communications* **12**, 53–75.

Mohammed, R., Rawashdeh, J. & Abdullah, M. (2020), Machine learning with oversampling and undersampling techniques: Overview study and experimental results, *in* '2020 11th International Conference on Information and Communication Systems (ICICS)', pp. 243–248.

Mosca, E., Szigeti, F., Tragianni, S., Gallagher, D. & Groh, G. (2022), SHAP-based explanation methods: A review for NLP interpretability, *in* N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond & S.-H. Na, eds, 'Proceedings of the 29th International Conference on Computational Linguistics', International Committee on Computational Linguistics, Gyeongju, Republic of Korea, pp. 4593–4603.
URL: *https://aclanthology.org/2022.coling-1.406/*

Rufibach, K. (2010), 'Use of brier score to assess binary predictions', *Journal of Clinical Epidemiology* **63**(8), 938–939.

Sadaiyandi, J., Arumugam, P., Sangaiah, A. K. & Zhang, C. (2023), 'Stratified sampling-based deep learning approach to increase prediction accuracy of unbalanced dataset', *Electronics* **12**(21).
URL: *https://www.mdpi.com/2079-9292/12/21/4423*

Shafer, G. & Vovk, V. (2008), 'A tutorial on conformal prediction', *Journal of Machine Learning Research* **9**(3).

Silva Filho, T., Song, H., Perello-Nieto, M. et al. (2023), 'Classifier calibration: A survey on how to assess and improve predicted class probabilities', *Machine Learning* **112**, 3211–3260.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J. & Baesens, B. (2012), 'New insights into churn prediction in the telecommunication sector: A profit driven data mining approach', *European Journal of Operational Research* **218**(1), 211–229.
URL: *https://www.sciencedirect.com/science/article/pii/S0377221711008599*

Vovk, V. & Petej, I. (2012), 'Venn-abers predictors', *arXiv preprint* .

Wongvorachan, T., He, S. & Bulut, O. (2023), 'A comparison of undersampling, oversampling, and smote methods for dealing with imbalanced classification in educational data mining', *Information* **14**(1), 54.

Xuan, S., Liu, G., Li, Z., Zheng, L., Wang, S. & Jiang, C. (2018), Random forest for credit card fraud detection, *in* '2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC)', IEEE, pp. 1–6.

Yan, G. (2024), 'Research on the application of alternative data in credit risk management', *Highlights in Business, Economics and Management* **40**, 1156–1160.

Zedda, S. (2024), 'Credit scoring: Does xgboost outperform logistic regression? a test on italian smes', *Research in International Business and Finance* **70**, 102397.

A APPENDICES