
# Evaluating the Impact of Machine Learning Models in SME Credit Risk Assessment

Eugeniu Litvinenco

Dissertation written under the supervision of Professor Nicolò Bertani

Dissertation submitted in partial fulfilment of requirements for the MSc in Business Analytics, at the Universidade Católica Portuguesa, January 2024.

# Evaluating the Impact of Machine Learning Models in SME Credit Risk Assessment

Eugeniu Litvinenco, January, 2024

**Abstract**

In recent years, there has been a significant increase in research efforts to incorporate machine learning (ML) models into credit risk assessment. This study focuses on the credit risk assessment of small and medium-sized enterprises (SMEs), which represent a significant source of employment in most economies. According to the regulator, while ML methods can provide added value, through a more accurate assessment of capital requirements, thus facilitating the access of this segment to financial services, there is a gap in the implementation of these methodologies in real-world context. This can be due to the overall complexity of explainability and interpretability, which forces financial institutions to use simpler models. Another reason is the lack of clarity on the benefits that these methodologies can provide. In this study, a hybrid model combining a decision tree and a logistic regression is proposed to address the complexity problem. This model shows comparable performance to the Random Forest and XGBoost while providing interpretability complexity equivalent to a logistic regression. In addition, this study introduces two innovative metrics, Exposure Weighted Distance to Default (EWDD) and Exposure Weighted Rating (EWR), which aim to find a way to distinguish the misclassifications made by a model according to their capital significance and to provide a sense of the total capital requirements that a model can generate. These metrics, along with the commonly used, were employed to compare the models, enabling the financial institutions to make a more informed decision in selecting the model that best meets their objectives.

**Keywords:** SME, machine learning, credit risk, hybrid model, exposure weighted metric

# Avaliação do Impacto de Modelos de Machine Learning na Avaliação do Risco de Crédito das PME

Eugeniu Litvinenco, Janeiro, 2024

**Resumo**

Recentemente, tem-se vindo a registar um aumento dos esforços de investigação na aplicação de modelos machine leraning (ML) na avaliação do risco de crédito. Este estudo centra-se na avaliação do risco de crédito das Pequenas e Médias Empresas (PME), que representam uma fonte significativa de emprego na maioria das economias. De acordo com o regulador, embora os métodos de ML possam proporcionar valor acrescentado, através de uma avaliação mais exata dos requisitos de capital facilitando assim o acesso deste segmento aos serviços financeiros, existe uma disparidade na implementação destas tecnologias no contexto real. Tal pode dever-se à complexidade da explicabilidade e interpretabilidade, que força as instituições financeiras a utilizar modelos mais simples. Outra razão é a falta de clareza sobre os benefícios que estas metodologias podem proporcionar. Neste estudo, é proposto um modelo híbrido que combina uma árvore de decisão e uma regressão logística contornando o problema da complexidade. Este modelo apresenta um desempenho comparável aos modelos Random Forest e XGBoost apresentando uma complexidade de interpretação equivalente à regressão logística. Adicionalmente, são propostas duas métricas, Exposure Weighted Distance to Default (EWDD) e Exposure Weighted Rating (EWR), cujo objetivo é distinguir as classificações erradas feitas por um modelo de acordo com a sua importância capital e fornecer uma perceção dos requisitos totais de capital que um modelo pode gerar. Estas métricas, juntamente com as métricas habituais, foram utilizadas para comparar os modelos, permitindo uma decisão mais informada na seleção do modelo que melhor satisfaz os objetivos de uma instituição financeira.

**Palavras-chave:** PME, machine learning, risco de crédito, modelo híbrido, métrica ponderada pela exposição

**Table of Contents**

**Acknowledgement**

I would like to express my gratitude to the entire Católica Lisbon community, including my supportive colleagues and friends, as well as the professors who sparked my interest in predictive modelling. Their dedication and enthusiasm have significantly influenced my academic and personal development.

I would like to particularly thank my supervisor, Nicolò Bertani, whose insightful guidance and encouragement of critical thinking has been essential in overcoming the challenges of my research.

I am grateful to the banking institution for providing access to necessary data and to its collaborators, whose cooperation and insights were important in shaping my thesis.

Lastly, I would like to express my heartfelt gratitude to my family, whose unwavering support has been my pillar. This achievement is dedicated to my parents, brother, and girlfriend, who have been by my side throughout this journey.

## 1. Introduction

Small and Medium Enterprises (SMEs) are crucial to the global economy, serving as a foundation for both developing and developed economies. In emerging economies, SMEs play a key role in driving innovation, reducing poverty, and stimulating economic growth. In developed economies, they make a significant contribution to employment, economic diversification, and overall economic stability. Accurate and reliable risk assessment not only supports SMEs access to financial resources but also promotes a healthier and more resilient financial ecosystem (Bańkowska et al., 2020). Effective credit risk management is essential in mitigating financial vulnerabilities and supporting the sustained growth and development of the SME segment, thereby strengthening the economy. In recent years, there has been a significant increase in attempts to incorporate machine learning (ML) models into credit risk assessment, with the purpose of enhancing the accuracy and efficiency of these evaluations, vital for this segment.

The European Banking Authority (EBA) has encouraged the use of ML techniques in credit risk assessment. However, there is still a significant gap in their practical implementation within the banking sector. This can be attributed to regulatory challenges, complexity of ML methodologies, particularly in terms of explainability, and a lack of clearness regarding its benefits for financial institutions. Explainability is essential in the banking industry as machine learning models often use complex internal mechanisms that are difficult to interpret. This opacity conflicts with the need for transparent and interpretable models that comply with regulatory standards and can be easily explained to regulators and stakeholders.

This study evaluates several ML techniques applied to the credit risk assessment of the SME segment and compares their performance with the operational, logistic regression based, model. The performance of the models is assessed in terms of their predictive power, measured by Area Under the Curve (AUC), and business impact, measured by exposure weighted metrics. The data used in this study was generously provided by a Portuguese bank and includes an ample historical record of the bank's SME portfolio over a five-year period, from October 2018 to October 2023.

The analysis begins with a Logistic Regression (LR) model and proceeding to K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Random Forest (RF), and eXtreme

Gradient Boosting (XGBoost) algorithms. Additionally, two hybrid models were developed. The first model is a hybrid of Decision Tree (DT) and Logistic Regression (DT+LR), combining their predictive capabilities in an interpretable model. The second hybrid model combines a Neural Network (NN) with a Logistic Regression, leveraging the NN's ability to capture complex patterns and the Logistic Regression's robustness (NN+LR).

The introduction of the Exposure Weighted Distance to Default (EWDD) and the Exposure Weighted Rating (EWR) metrics is motivated by the recognition that better predictive performance in machine learning models may not always align with the best interests of banking institutions. Traditional metrics as AUC treat all misclassifications equally, which is an oversimplification in the complex landscape of credit risk assessment. In practice, misclassifying an instance with a higher outstanding loan is more consequential than an error on a smaller loan. Therefore, a more nuanced and financially sensitive approach to model evaluation is necessary. The EWDD provides a detailed understanding of the underestimation errors in the probability of default, weighted by the financial exposure of each loan, addressing this concern. This ensures that the performance is evaluated with a focus on the errors that present the greatest financial significance, prioritising the economic stability. At the same time, the EWR provides a measure of the overall conservatism of the model, indicating its tendency to overestimate risk. While a conservative model may safeguard against risk, it could also result in missed opportunities and reduced profitability.

The study highlights the importance of considering not only the predictive performance but also the business impact of models in credit risk assessment. The findings indicate that incorporating exposure weighted metrics can lead to more insightful decisions regarding model selection, particularly in a real-world business context highlighting the benefits of ML techniques. Moreover, in terms of the proposed metrics, the hybrid model DT+LR exhibited promising results, which are comparable to those of the ensemble models. This outcome suggests that hybrid models can be a viable option, providing a balance between performance and interpretability, which is essential for practical applications in financial institutions.

The following chapter provides an overview of credit risk assessment in the SME segment, followed by a literature review of related research. The subsequent chapter briefly reviews the methodology of the models, with a detailed explanation of the development of the hybrid models and the proposed metrics. The ensuing chapter presents the data, along with the related challenges and assumptions. Finally, the last chapters present the results obtained for each

model and metric, followed by a discussion of these results, including limitations and suggestions for future research.

The files containing the complete process, including data wrangling, model development, and model comparison, can be found at: https://github.com/EugeniuLitvinenco/thesis/tree/09b09a1b4c48fd14977248365e4a94a7ae5def16

## 2. Background

This chapter examines the evolving landscape of credit risk assessment in the Small and Medium Enterprises (SMEs) sector, which is a crucial component of global economic stability and growth. The regulatory context is outlined, highlighting the significance of SMEs and the various efforts being made to improve their access to capital. Additionally, it discusses the transition from traditional credit risk models to the incorporation of machine learning techniques in response to the increasing demand for more nuanced assessment methods. It identifies the main difficulties of this transition and explores the main approaches designed to overcome them.

### 2.1. Regulatory Context

The importance of Small and Medium Enterprises (SMEs) in modern economy has been aim of research across the world. In developing countries, this segment is seen as a poverty reducer and a driver of development and growth. In developed countries, it is considered to add robustness and stability to the growth of the economy (Abdulsaleh & Worthington, 2013). Although there is no conclusive research proving a causal effect of SME development on economic growth, this sector represents a significant source of employment in most of economies. Furthermore, the Organisation for Economic Co-operation and Development (OECD) emphasised the importance of SMEs in the OECD Green Growth Strategy in 2012, highlighting that the transition of this sector is crucial to the overall greening of the economy (Eugene Mazur, 2012). This report anticipates the understanding of the European Investment Fund that, on its working paper (Kraemer-Eis et al., 2012), exploring the access to capital of this segment, states that the most widespread debt financing methods are bank loans and overdrafts. Limited access to finance remains a significant constraint that SMEs face according

to (Rao et al., 2023). This systematic literature review provides an overview of the research conducted in this field, highlighting the emerging innovative methods of capital raising, such as crowdfunding and ICOs (Initial Coin Offerings - a type of funding using cryptocurrencies). However, the conventional sources of finance, bank loans, represent the majority.

In March 2016, the European Bank Authority (EBA) introduced a capital reduction factor for loans to SMEs. The EBA stated that SMEs are key actors in the European Union (EU) economy in terms of their share in employment and added value. This measure counterbalances the rise in capital requirements resulting from the Capital Conservation Buffer, thereby stimulating the credit flow for this particular segment (EBA, 2016).

To homogenise the risk assessment of credit risk in the EU, the European Central Bank (ECB) has proposed the Internal Rating Based (IRB) approach. This approach allows banks to use internal models to estimate the Probability of Default of retail, corporate, and sovereign borrowers. This approach enables banks to estimate their own capital requirements in compliance with the Capital Requirements Regulation (CRR), ensuring that they hold an appropriate amount of capital to cover potential losses.

Credit risk assessment can typically be divided into two types: concession and monitoring. Concession risk assessment pertains to the initial decision-making process, during which a financial institution evaluates whether to grant a loan to a prospective borrower. The focus of this study is on monitoring risk assessment within the SME segment. Monitoring is a continuous process, often on a monthly basis, in which financial institutions estimate the probability of a borrower defaulting on its loan obligations within the forthcoming 12 months. In other words, monitoring risk assessment involves the ongoing evaluation of the creditworthiness of borrowers.

Figure 1, presents an overview of the credit risk monitoring process, using internal models, applied to the SME segment. The enterprise, in this case classified as Small or Medium according to EBA guidelines (EBA, 2016), generates financial data. This data is then used by the financial institution to estimate the probability of the enterprise defaulting within 1 year. The resulting probability is then mapped onto a standardized rating scale, which typically ranges from 1 (low risk) to 9 (default). The position of the enterprise on this scale provides a measure of credit risk. This measure is used to estimate the capital requirements and when aggregated across the portfolio enables the bank to assess the overall risk exposure ensuring that risk management strategies reflect the current economic realities.

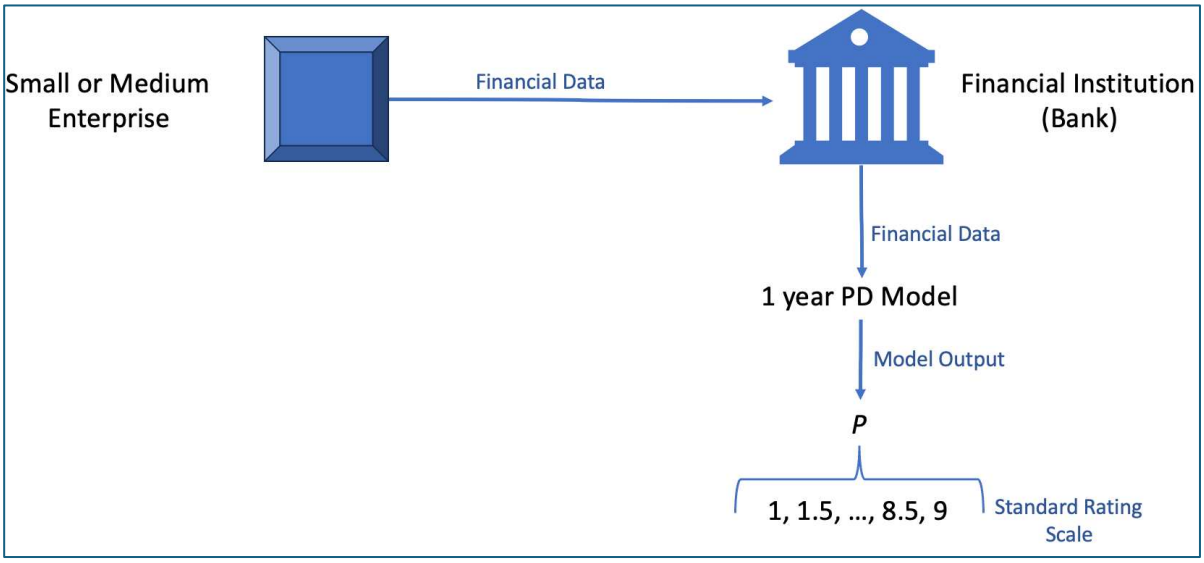


*Figure 1 – High-level overview of Credit Risk Assessment on SME segment*

According to (ECB, 2023) a candidate for the IRB approach must meet specific minimum requirements, including model documentation. This documentation should enable a third party to independently understand the methodologies, assumptions and limitations, and replicate the model development - implementation process. Only financial institutions approved by the national supervisor are permitted to use internal models to estimate their capital requirements. From the bank's perspective, this has been perceived as an opportunity to make accurate capital provisioning estimates that avoid excessive overestimation and still meet requirements. This involves the use of more empowered predictive models that can better distinguish between "bad" and "good" borrowers. Therefore, this opportunity combined with exponential increase in data storing capacity coupled with improvements in computing power, resulted in a significant effort in applying Machine Learning models in credit risk assessment.

According to (ECB, 2023) the probability of default (PD) models should be transparent i.e. the model should be interpretable and explainable. The EBA has published a Discussion Paper on Machine Learning for IRB Models (EBA, 2021) which recognises the improved predictive power of ML-based internal models. However, it also highlights that the incorporation of these techniques into institutions' IRB approach has not been as rapid as in other areas. The document states that ML models can provide added value, but they should comply with CRR. It anticipates the three main obstacles that financial institutions have faced during the transition process. The primary issues arise from the complexity of the results, making it challenging to interpret and explain them to management and supervisors. In (EBA, 2020) interpretability of a model is defined as the ability of humans to directly understand its internal behaviour, while

explainability referred to the ability to provide justifications for the main factors that led to its output.

### 2.2. Credit Risk Assessment and Explainability

Credit Risk models have evolved significantly in the last century. Edward I. Altman on (Altman, 2018) celebrates the 50th anniversary of his PhD graduation, (Altman, 1968), with a retrospective of the evolution of the credit risk methodologies, including his own Altman Z-Score. The main corporate scoring systems over time stated by Altman are the following:

- Qualitative models – 1800s
- Univariate analysis - early 1900s
- Multivariate analysis (Z-score) – late 1960s to the present:
    - discriminant, logit and probit models
    - nonlinear and "black box" models, neural networks (1990s)
- Artificial intelligence systems – late 1990s to the present:
    - expert systems
    - neural networks
    - machine learning

(Altman, 2018) points out, multiple types of models estimating the Probability of Default coexist simultaneously.

The implementation of ML models in credit risk assessment have been a notorious research subject. According to (Bhatore et al., 2020) ensemble techniques have outperformed single classifiers in several studies. In 2022, a systematic literature review of 76 papers (Shi et al., 2022) identified and discussed the major challenges in ML-driven credit risk assessment and ranked the models according to their performance based on public datasets. The major challenges and incentives for further research were related to data imbalance, fortunately there are more observations of non-defaulting firms than defaulting ones, inconsistency of datasets, model transparency, and inadequate use of deep learning models. The review concludes that deep learning models generally outperform other models, and ensemble methods tend to provide higher accuracy compared to single models. A more recent review (Noriega et al., 2023) points out some of the same challenges and emphasizes the transparency constraint in credit risk. Furthermore, this study recognizes the Boosted category as the most extensively researched family within ML methodologies. It also highlights that the most commonly used metrics are the Area Under the Curve (AUC) and Accuracy (ACC).

The dominance in performance of ensemble and boosting techniques, not exclusively in credit risk assessment, have been compromised by their complexity and concerns regarding explainability and interpretability. The latter has been a subject of research in both financial and medical sectors (Ribeiro et al., 2016) (S. M. Lundberg et al., 2020).

XAI stands for eXplainable Artificial Intelligence and according to (Gunning et al., 2019) its purpose is to make the behaviour of AI and ML models understandable for humans by providing explanations. In (Barredo Arrieta et al., 2020) the authors introduce the concept of responsible AI, which includes fairness, model explainability and accountability as its core. The shift of the paradigm from performance-centric to explainability-centric, where explainability of the model can be even more important factor than its performance is well described in (Burkart & Huber, 2021) where the eight main aspects motivating explainability are stated:

- **Trust**: Understanding the model's strengths and weaknesses.
- **Causality**: Explainability, through attribute importance, provides a sense of causality.
- **Transferability**: A predictive model must convey its future behaviour to be used with unseen data.
- **Informativeness**: A system must serve its real-world purposes.
- **Fair and Ethical Decision Making**: Decision-makers must present comprehensible results to ensure commitment to ethical standards.
- **Accountability**: An algorithm must be able to explain and justify its decisions to be accountable.
- **Adjustability:** Explainability allows domain experts to adjust the prediction model by incorporating domain knowledge.
- **Proxy Functionality:** Explainability serves as a proxy when examined based on criteria that cannot be easily quantified.

Several significant works proposing methods for local explainability have been published in the last decade. This involves calculating the sequence of node decisions that led to a prediction, given an instance (Delgado-Panadero et al., 2022; S. M. Lundberg et al., 2019, 2020; Ribeiro et al., 2016; S. Lundberg & Lee, 2017).

### 2.2.1. Explainability Approaches

The metareview of XAI (Clement et al., 2023) illustrates in Figure 2 the current research objectives, which aim to achieve the high complexity and high explainability region.

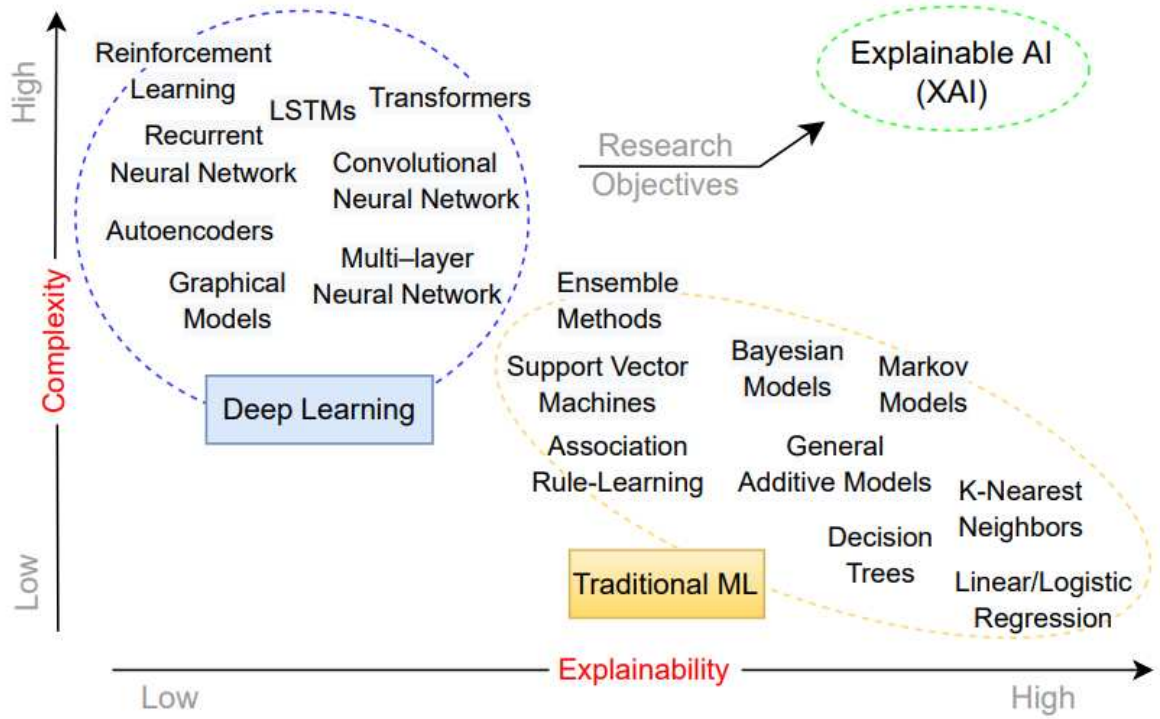


*Figure 2- XAI Complexity-Explainability, from* (Clement et al., 2023)

One impactful methodology known as Model-Agnostic approach was described (Baehrens et al., 2010) and (Ribeiro et al., 2016). These approach treats the Machine Learning models as black-box functions.

### 2.2.2. Local Interpretable Model-Agnostic Explanations (LIME)

(Ribeiro et al., 2016) proposed Local Interpretable Model-Agnostic Explanations (LIME), explanation approach of significant contribution on the ongoing discussion on interpretability of ML models.

According to (Ribeiro et al., 2016), LIME involves the following steps:

- Data Perturbation: A new dataset is generated applying perturbations to samples and corresponding predictions.
- Auxiliar model: An interpretable model is trained and weighted by the proximity to the instance of interest.

- Local Fidelity: The interpretable model represents an approximation of the original model predictions locally.

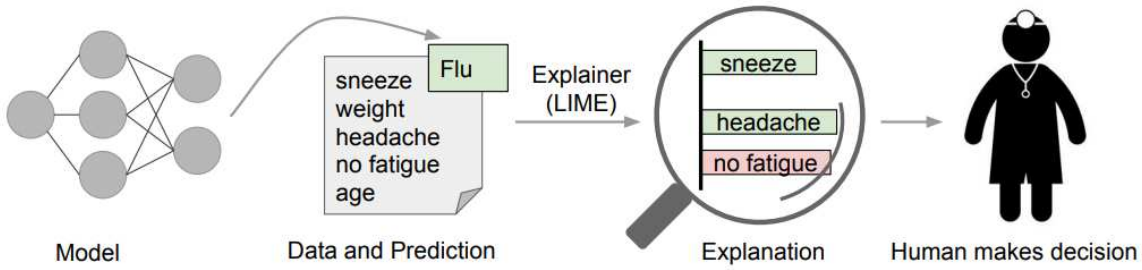


*Figure 3- LIME explaining individual predictions,* (Ribeiro et al., 2016)

### 2.2.3. SHapley Additive exPlanations (SHAP)

Another approach to this problem is the SHAP framework. First introduced in (Lloyd Shapley, 1953) and awarded with a Nobel Memorial Prize in Economic Sciences in 2012, Shapley Values represent a solution concept in cooperative game theory that turned out to be particularly useful in context of ML. (S. Lundberg & Lee, 2017) presents a unified framework for interpreting predictions, SHAP (SHapley Additive exPlanations) and (Rozemberczki et al., 2022) were among the first to emphasize the use of Shapley Values in the context of machine learning. According to (S. Lundberg & Lee, 2017), SHAP assigns each feature an importance value given a particular prediction.

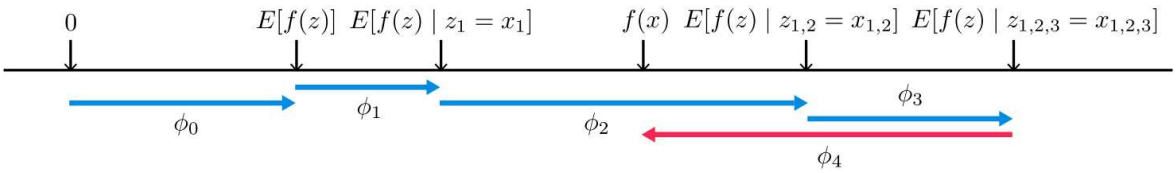


*Figure 4 – SHAP values and the expected feature importance, (S. Lundberg & Lee, 2017)*

The main advantage of LIME and SHAP is being Model-Agnostic which makes them suitable for any classifier. However, they provide local approximations with local fidelity, which may not always give a complete picture of the model's behaviour. According to the recent paper (Salih et al., 2023), LIME and SHAP, widely used specially with tabular data, have advantages and issues and the choice may vary depending on the specific task.

### 2.3.4. Other Relevant studies

(Altman & Sabato, 2005) conducted a comparative study between a traditional generic Multivariate Discriminant Analysis (MDA) model and a Logistic Regression. They found that the Logistic Regression outperformed the MDA model by almost 30% in a held-out sample. Another approach was taken in (Roy & Shaw, 2021), the researchers developed a credit scoring model for SMEs combining the Best Worst Method (BWM) and the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) creating a multicriteria model to guide the decision-making process. The resulting multicriteria model was found to be more accurate than an existing commercial model, despite the absence of more sophisticated models.

According to (Bhatore et al., 2020; Noriega et al., 2023; Shi et al., 2022) the most explored approach consists of constructing complex ML models and applying previously discussed SHAP or LIME techniques to explain their predictions. (Bussmann et al., 2021a) in a sample of 15,000 small and medium enterprises used the SHAP framework to group the borrowers according to their financial characteristics and explain their credit score and to predict their future behaviour. The researchers compared the Area Under the ROC Curve of the best configurations of a Logistic Regression and a XGBoost model observing an increase from 0.81 to 0.93. The authors emphasize that the lack of interpretability and auditability of AI and some ML methods represents a macro-level risk. They state that even though these methods can improve credit risk assessment, and thus the accessibility of financial services, it can also introduce new risks.

Similarly, (de Lange et al., 2022) employed the SHAP framework on a real dataset. The researchers combined Light Gradient Boosting Machine (LightGBM) with SHAP, to interpret which and how the explanatory variables affect the predictions. This approach outperforms the used in real context model based on a Logistic Regression. The authors clearly illustrate the influence of distinct variables on the predictions, finding that the volatility of utilized credit balance and the duration of the customer relationship were the most important explanatory variables. The authors used the models to estimate the Loss Given Default (LGD) and suggested this as a method for analysing the potential economic benefits of an improved credit scoring model.

Regarding the current situation of ML models in the real-world credit assessment, according to the Follow-Up report on the discussion on ML for IRB Models (EBA, 2023), most financial institutions use variants of regression analysis, while fewer use more advanced ML techniques.

One main reason is the overall complexity, development and explainability, which can lead to inability of the institution to accurately document and justify the model and its predictions, leading to rejection by regulators. In order to address this reason this study explores the construction of a hybrid model of relatively low interpretability and explainability complexity aiming to achieve a reasonable complexity-performance trade-off. Therefore, this study proposes an alternative approach to model development which pretends to benefit from both, Decision Tree and Logistic Regression models, combining them. This approach differs from the work presented in (Sagi & Rokach, 2021), where a Decision Tree that approximates the predictive performance of a pre-trained ensemble of trees is built. In contrast, the presented approach involves building a Decision Tree from which rules are extracted to create new features. These new features, along with the original features, are then used to train a Logistic Ridge Regression. The interpretability and explainability of this hybrid model remain comparable to that of Logistic Regression.

Another potential inhibitor of the transition to ML models can be the managerial underestimation of its benefits. This study addresses this factor designing additional metrics, Exposure Weighted Distance to Default (EWDD) and Exposure Weighted Rating (EWR), that aim to provide an indirect estimation of the capital requirements the models could generate, thus providing a sense of the benefits.

The developed models in this study are compared to the existing, operational, Logistic Regression based model, according to the commonly used and the proposed metrics.

## 3. Methodology

This chapter outlines the methodology used to assess credit risk in Small and Medium Enterprises (SMEs), with a focus on data preparation, model selection, and performance evaluation. The chapter begins with the important topic of data splitting, where a custom methodology is presented to enhance the validity of model scores. This is followed by an explanation of the pre-processing steps taken to standardize and optimize the data for machine learning applications. The chapter reviews the predictive models used in this study, including Logistic Regression with Ridge penalization, Random Forest, and XGBoost,

chosen for their relevance in credit risk assessment. It also explores innovative stacking techniques and the calibration process necessary for accurate probability estimations. Finally, two additional metrics are introduced, Exposure Weighted Distance to Default (EWDD) and Exposure Weighted Rating (EWR), designed to provide a deeper understanding of the economic and risk management implications of each model.

### 3.1. Splitting

The process of data splitting is fundamental for the training and validation of models, especially when evaluating their performance. The division of data into training and testing sets is essential to assess how effectively a model generalizes to new, unseen data, providing a robust evaluation of its performance.

In credit risk assessment, traditional random sampling for data splitting, which involves selecting rows at random, can lead to an overestimation of performance metrics due to interpolation. This issue arises when observations related to the default status of an enterprise in distinct dates are present in both training and testing sets. For instance, if the testing set includes the Target variable for an enterprise in month $n$, while the training set includes the Target variable for the same enterprise in month $n-1$ and $n+1$, this allocation may result in an underestimation of the testing error, making it easier for the model to predict. As a consequence, the assessment of the overall performance of the model may be compromised.

To prevent this interpolation problem, a custom splitting function was designed. Rather than employing row-wise sampling, this function randomly assigns enterprise identifiers to the training and testing sets while maintaining the proportion of ever defaulted firms in each set. It is essential to maintain this proportion in both the training and test datasets to avoid bias, ensure accurate performance estimation and maintain the predictive reliability of the models in real-world risk assessments.

This method was applied consistently throughout all stages of data splitting, including the creation of folds for the cross-validation process.

### 3.2. Pre-processing

Pre-processing steps for model training were extensive, including standardisation of numerical features and encoding of categorical variables to ensure uniformity and compatibility with machine learning algorithms. Undersampling was applied to balance the dataset and potentially

improve the predictive power of the models. An important aspect of the pre-processing was the removal of highly correlated features. This crucial step reduced multicollinearity, allowing the models to focus on independent and relevant predictors. This careful preparation ensures that the models remain robust and reliable in their predictions.

### 3.3. Brief Review of the Main Predictive Models Used in This Study

#### 3.3.1. Logistic Regression with Ridge penalization

Logistic Regression (LR) is a binary classification algorithm commonly used in machine learning. It models the probability of an instance belonging to one of two classes, in this case, "bad" or "good" borrower. The model employs a linear combination of input features, that are transformed through a sigmoid function. The optimization is guided by the binary cross-entropy cost function.

The probability of instance *(Y, X)* belonging to class *1* is given by:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

The optimization problem focuses on minimizing the cost function which quantifies the discrepancy between the predicted probabilities and the actual class labels. This guides the model to adjust its parameters for more accurate predictions.

$$\text{Cost}(y, \hat{y}) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)} \log\left(\widehat{y^{(i)}}\right) + \left(1 - y^{(i)}\right) \log\left(1 - \widehat{y^{(i)}}\right)\right]$$

In machine learning, overfitting can occur when the model starts interpreting noise in the data as information. To prevent this, a regularization technique should be applied. Logistic Ridge Regression is a modification of Logistic Regression that incorporates Ridge regularization by adding a penalty term to the cost function. This penalizes large coefficients, preventing the model from assigning excessively large weights to features and controlling the bias-variance trade-off.

$$\text{Cost}_{\text{Ridge}}(y, \hat{y}) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)} \log\left(\widehat{y^{(i)}}\right) + \left(1 - y^{(i)}\right) \log\left(1 - \widehat{y^{(i)}}\right)\right] + \frac{1}{C}\sum_{j=1}^{n} \beta_j^2$$

The regularization strength is controlled by the hyperparameter *C*, which is the inverse of the regularization strength. As *C* decreases, the flexibility of the regression fit decreases, resulting in a reduction in variance but an increase in bias. This balance enables better generalization to unseen data (James et al., 2023b). Therefore, finding the optimal hyperparameter *C* allows for better generalization to unseen data.

Logistic Regression is one of the most used models in credit risk assessment. These models are highly explainable (Clement et al., 2023), providing coefficients for each feature allowing for a clear understanding of how each feature influences the prediction.

A tuning procedure was used to approximate the optimal C, which, using linear relationships, extracts the most information from the training data while avoiding overfitting. Random values between $1.0 \times 10^{-5}$ and $0.5$ were tested for the hyperparameter C, and the performance of the corresponding model was evaluated on an unseen subset of the data during the Cross-Validation procedure. The final predictions were made using the configuration of the model with the highest area under the ROC.

### 3.3.2. Random Forest

A random forest consists of a combination of tree predictors, where each tree depends on the values of a random vector that is sampled independently and with the same distribution for all trees in the forest (Breiman, 2001). The random forest is an ensembled model that uses bagging. Multiple decision trees are employed in parallel, with each tree being provided with a subset of the data through bootstrapping. The predictions of this trees are combined to produce a more accurate and robust result. Additionally, random subsets of features are considered at each split to prevent the decision trees from becoming specialized only on particular features. In classification problems, predictions are made through a voting mechanism leading to accurate and stable result. Random forest classifiers are generally robust to overfitting since the number of trees is sufficiently high (James et al., 2023e).

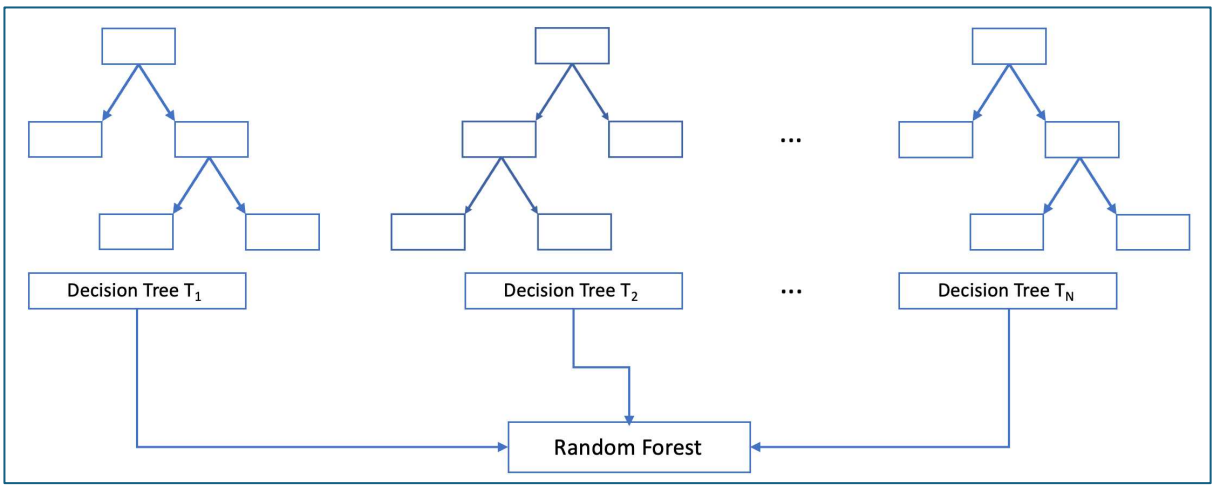


*Figure 5 - Random Forest Structure*

In this study, two hyperparameters of the Random Forest model where tuned:

- *n_estimators*, the number of trees combined in the forest.
- *max_depth*, maximum depth of each tree.

The number of trees in the forest was selected at random from a range of 100 to 200.
As pointed out by (Cutler et al., 2012) when the model was introduced in (Breiman, 2001), it was recommended to use large values for the maximum depth of each tree. However, (Segal & Xiao, 2011) demonstrated that this practice does not necessarily lead to the optimal solution. Therefore, the maximum depth was randomly selected from a range of 3 to 5.

### 3.3.3. XGBoost

XGBoost, eXtreme Gradient Boosting, is an optimized distributed Gradient Boosting library designed to be highly efficient, adaptable, and portable. It is widely used by data scientists and provides state-of-the-art results on many problems (Chen & Guestrin, 2016). Similar to Random Forest, Gradient Boosting Decision Trees is an ensemble learning method that constructs multiple decision trees independently using subsets of the training data. The predictions are generated by calculating a weighted sum of the predictions made by each individual decision trees.
Boosting is an alternative method for improving the predictions of a decision tree based model. Unlike Random Forest, Boosting develops a sequence of trees, each using information from the previous trees. Bootstrap sampling is not involved in Boosting, and the trees are fitted using residuals from the previous trees (James et al., 2023f).

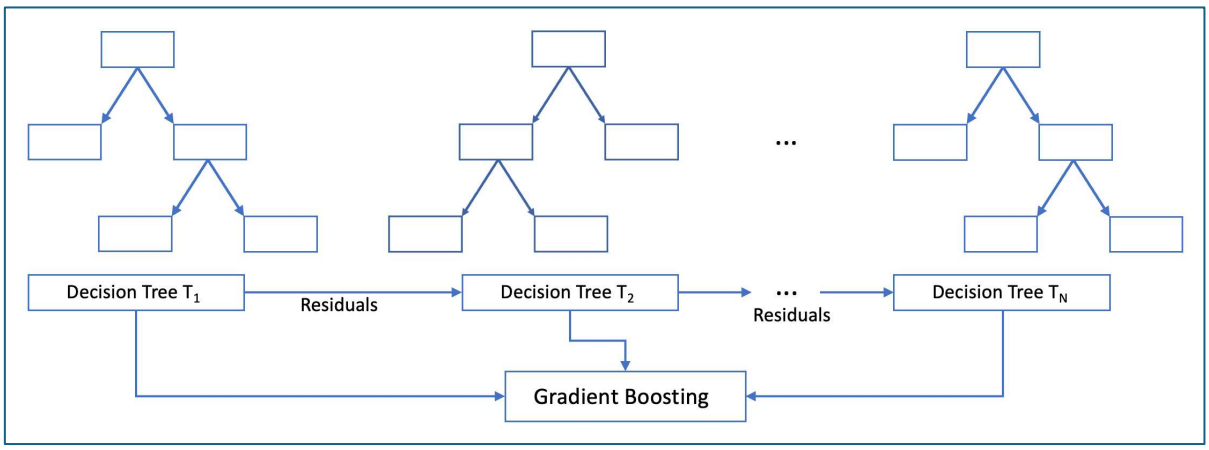


*Figure 6 - Gradient Boosting Structure*

During the tuning procedure three hyperparameters were estimated:

- λ, the shrinkage parameter also known as the *learning_rate,* controls the rate at which the boosting algorithm learns.
- *n_estimators*, the number of trees.
- *max_depth*, maximum depth of each tree.

In contrast to Random Forest, Boosting can potentially start overfitting if the number of estimators is too large. Considering the recommendations of (James et al., 2023f), when tuning the hyperparameters the number of estimators was randomly selected between 100 and 200. Understanding that this can lead to overfitting a low learning rate interval was tested, with a uniform distribution between $1.0 \times 10^{-5}$ and $1.0 \times 10^{-4}$. The maximum depth of the trees was kept within an interval of 3 to 5.

This algorithm has faced criticism for its bias towards features with a large number of potential splits, which can lead to interpretability issues and overfitting constraints. Recent research (Zhang et al., 2023) has revealed that these issues stem from two sources: a systematic bias in the gain estimation for each split, and a bias in the split-finding algorithm. This bias arises from using the same dataset to both assess split improvement and identify the optimal split. The use of cross-validation ensures that the model is tested on different subsets of the data, thereby mitigating the potential effects of bias.

### 3.3.4. Additional Models

Additionally, out of curiosity, two classifiers were tested for their performance on this particular data, SVC and KNN. The Support Vector Classifier (SVC) estimates a hyperplane to separate classes and makes predictions based on the side on which the observation falls. The SVC can be written as an optimization problem where the objective of the algorithm is to find the hyperplane that maximizes the margin between the classes. SVC may misclassify some observations and the severity of this misclassification can be specified and tuned (James et al., 2023d). The K-Nearest Neighbors (KNN) classifier identifies the K observations that are closest to the one being predicted and estimates the conditional probability of it being of the same class as the others (James et al., 2023c). The algorithm considers either Euclidian or Manhattan distance to assess the closeness between observations, and the hyperparameter K should be an integer.

### 3.3.5. Stacking

#### 3.3.5.1. Stacking Decision Tree and Logistic Regression

To further improve the predictive capabilities of the logistic regression, a stacking technique was applied. This technique aims to leverage the predictive advantages of both regression and tree models, enriching the feature space of the model by integrating decision tree rules.

The Decision Tree model is a fundamental principle used in various other complex models. It consists of an ordered and structured set of decision rules that minimize heterogeneity within each group that satisfies the rules. In a Decision Tree, each internal node represents a feature, each branch represents a decision rule, and each leaf node represents an outcome. The algorithm partitions the data based on feature values using a recursive partitioning method (Breiman et al., 1984).

In essence, a Decision Tree model involves recursively dividing the training data based on the values of distinct features. The features are chosen using a criterion such as Gini or Entropy Impurity functions (James et al., 2023f).

$$\text{Gini} = \sum_{k=1}^{K} \widehat{p_{mk}}(1 - \widehat{p_{mk}})$$

Where $\widehat{p_{mk}}$ is the proportion of training observations in region *m, K* are the classes. A small Gini Index indicates that region *m* contains predominantly observations from the same class.

$$\text{Entropy} = -\sum_{k=1}^{K} \widehat{p_{mk}} \log(\widehat{p_{mk}})$$

The Entropy on region *m* is zero if all instances of the nodes are from a single class.

The partitioning process continues until at least one of the following conditions is satisfied:

- The maximum depth is reached.
- It is not possible to continue partitioning respecting the minimum number of instances in each leaf node.

These models are easily interpretable which makes them particularly interesting for credit risk purposes.

According to (Zhao & Nie, 2021), the Entropy criterion can be more influenced by noise variables and missing values in the data, leading to bias. Nevertheless, during the tuning procedure, both Gini and Entropy criterions were considered.

Additionally, the depth and the minimum number of samples required to split an internal node of the Decision Tree were tuned. The depth was randomly selected between 4 and 5, this range was chosen based on the understanding that a tree with a large depth might overfit the data, while a tree with a small depth might not capture the important patterns, i.e. underfit. Regarding the minimum number of samples required to split an internal node, the parameter was chosen from a uniform distribution between 0.10 and 0.90, to ensure that enough instances are present in a node before it is split. The minimum number of samples in each leaf was set at 10.

Figure 7 illustrates the workflow, it starts with the partitioning of the training dataset, continues with the training of the decision tree, proceeds through the extraction and translation of the tree rules into binary features, and concludes with the training of the logistic regression model on the extended dataset.

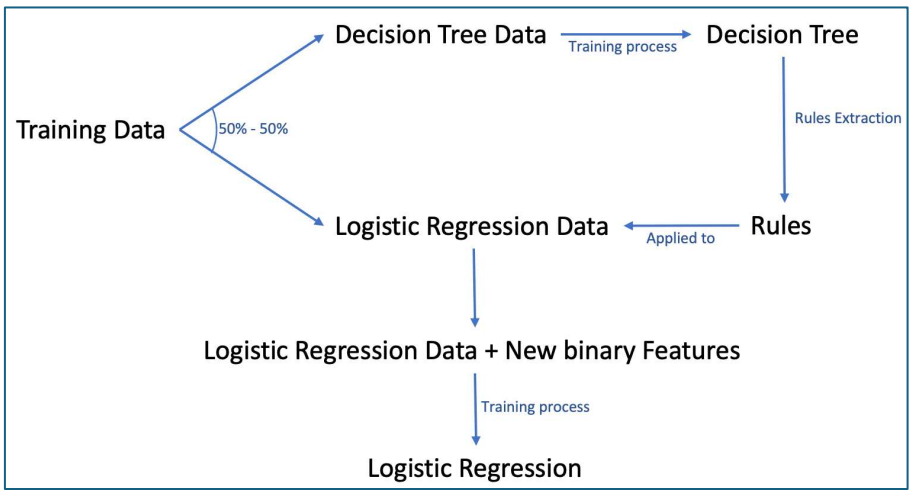


*Figure 7 - Decision Tree and Logistic Regression Stacking*

One of the key benefits of this stacked model is its interpretability, which is particularly valuable in credit risk analysis where model transparency is critical. The explicit inclusion of rule-based features from the decision tree allows for a greater degree of model interpretability compared to more opaque ensemble methods. A decision rule is a set of conditions derived from the structure of the decision tree that determines the path data takes to reach a particular leaf node. Each rule represents a particular combination of feature values that leads to a decision. Figure 8 presents an example of a possible tree structure. In order to achieve the *Leaf n,* instances should satisfy the following condition: If the *Turnover_AMT* is lower or equal to 500,000 and the *Resp_Banca_VS_EBITDA* is lower or equal to 10, then *Leaf n.*

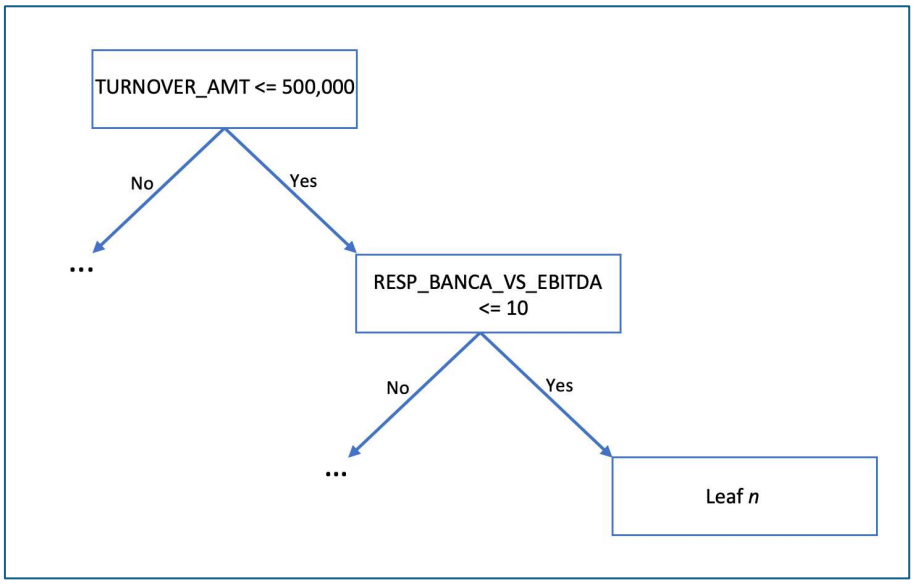


*Figure 8 - Example of a Decision Rule*

These rules are extracted from the DT model and then applied to unseen data to create new features. Each of these rules is interpretable and is stored with the original data in the form of binary variables.

This feature augmentation is expected to either improve or at least maintain the predictive performance achieved by the logistic regression model in isolation. The integration of these decision tree rules adds a layer of complexity to the feature set, enabling the detection of intricate patterns and interactions that a stand-alone logistic regression model may not be able to capture.

#### 3.3.5.2. Stacking NN + LR

In order to more effectively approximate the upper limit of performance achievable through feature engineering with this dataset, a second stacked model was developed. In this hybrid model, a neural network (NN) is first trained on a dedicated subset of the training data. The trained NN is then used to make predictions on the remaining training data, where the features extracted from the last hidden layer of the NN are used to train a logistic regression model. The workflow of this stacking procedure is shown in Figure 9.

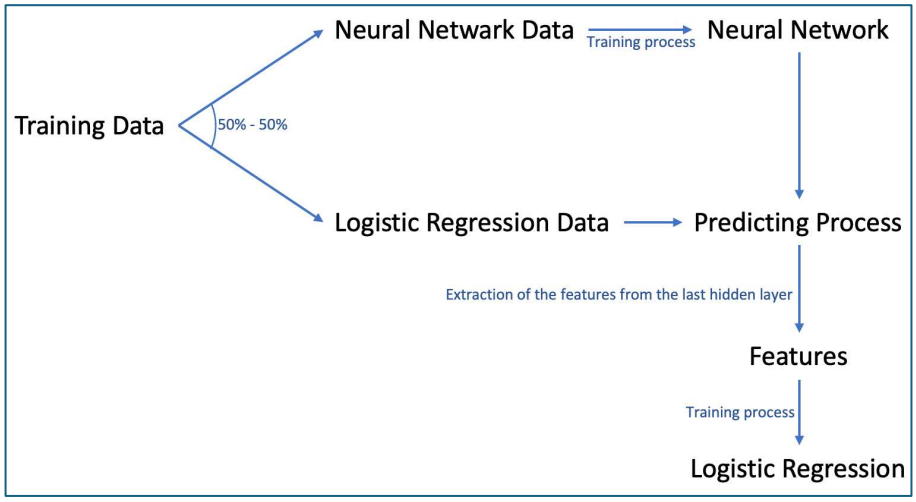


*Figure 9 - Neural Network and Logistic Regression stacking process*

This approach was strategically designed to exploit the ability of NNs to capture complex patterns in order to extend the feature space beyond what is accessible by traditional feature engineering techniques. It is based on the Universal Approximation Theorem (Cybenko, 1989), which asserts the ability of NNs to approximate a wide range of functions. This study assumes the potential of NNs to identify and extract the most predictive empowered set of features from

the original dataset. Thus, it is hypothesised that the performance of the stacked model, which combines an NN with an LR, could potentially approximate the performance of a LR model trained on an optimally designed feature set.

### 3.4. Calibration

In order to convert the predictions of the models into probabilities all predictions were calibrated. Calibration is a crucial process in probability estimation using ML techniques. As described in (Silva Filho et al., 2023; Van Calster et al., 2019) this process is essential for accurate probability estimation, especially in imbalanced datasets or when using models that are inherently weak in probability estimation. The predictions of all models developed in this study were refined using the *CalibratedClassifierCV* method, with the incorporated *sigmoid* function. The *sigmoid* function smoothly maps the predictions of the models to probabilities between 0 and 1. This non-linear transformation is particularly adept at handling extreme predictions, making it appropriate for probability calibration in the models used in this study.

### 3.5. Metrics

#### 3.5.1. Usual Metrics

According to (Noriega et al., 2023; Shi et al., 2022), the most commonly used metrics for assessing model performance in credit risk are the ROC AUC (Receiver Operating Characteristic Area Under the Curve) and ACC (Accuracy). In line with this established practice, this study also employs these metrics for performance assessment.

#### 3.5.2. Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC)

The ROC curve is a graphical representation that illustrates the diagnostic ability of a binary classifier as its discrimination threshold is varied showing two types of errors, True Positive Rate (TPR), also known as sensitivity, and False Positive Rate (FPR) (James et al., 2023a). For each decision threshold the ROC curve results in points (1-FPR, TPR), 1-FPR is also known as specificity.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}};\ \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}},$$

$$where\ TP = sum\ of\ true\ positive;\ FN = sum\ of\ false\ negative;\ TN = sum\ of\ true\ negative$$

The area under the ROC curve provides a measure of the discriminatory power of the model, translating how well a model can distinguish a “good” from a “bad” borrower. A model with perfect predictions has an AUC of 1, while an AUC of 0.5 represents randomness.

### 3.5.3. Accuracy (ACC)

The accuracy is defined as the proportion of true predictions (both true positives and true negatives) in the total population.

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}},$$

$$where\ TP = sum\ of\ true\ positive;\ FN = sum\ of\ false\ negative;\ TN = sum\ of\ true\ negative$$

This metric is not the most representative of model performance, especially in imbalanced datasets, which is the case. For example, in a highly imbalanced dataset where the positive class represents 1% of the total observations, even the model that always predicts the negative class has an ACC of 99%.

### 3.5.4. Additional Metrics

One of the limitations of the existing metrics is the inability to observe the score of an enterprise between 1 and 8.5. In fact, the only observable score is 9, which represents default.
Indirectly, researchers mainly assess the discriminative power of a model through ROC AUC, ACC, F1 (Bhatore et al., 2020; Bussmann et al., 2021a, 2021b; de Lange et al., 2022; Noriega et al., 2023; Sagi & Rokach, 2021; Shi et al., 2022).
However, since these models are designed to estimate 1-year PD, it would be informative and instructive when comparing models to have an estimate of how far off the model's predictions are, on average, *n* months before enterprises default.
When assessing the credit risk of a portfolio of borrowers, two different types of error can occur:

- Underestimation of the PD, implying lower capital retention which can potentially lead to bankruptcy of the financial institution.
- Overestimation of the PD, implying excessive capital retention leading to loss of potential added value.

The main focus of a credit risk model should be minimizing the underestimation of the PD not allowing excessive overestimation errors.
As a form of model comparison, this study proposes the following metrics.

### 3.5.5. Exposure Weighted Distance to Default (EWDD)

In order to assess the underestimation errors of a model when estimating 1 year probability of default of a portfolio of borrowers, the following metric is proposed:

$$EWDD_n = \frac{\sum_d \left(\text{EXPOSURE}_{d,t} \times \left(9 - \text{RATING}_{d,t-n}\right)\right)}{\sum_d \left(\text{EXPOSURE}_{d,t}\right)}, \forall \text{d in P},$$

where *P* is the entire portfolio, *d* represents the defaulted enterprises, *t* represents the default month of the enterprise *d,* and *n* is the number of months before the default.
In this formula, the predicted rating of a defaulted enterprise *n* months before default is compared to the rating of an enterprise in default, *9*. The lower the *n,* the smaller should be this difference. Additionally, in order to penalize higher capital significance, the distance to default, in the rating scale, is weighted by the exposure of the enterprise on the default month. Exposure in this formula represents the outstanding loan capital. The denominator serves as a scaling factor allowing a more comprehensive interpretation.

### 3.5.6. Exposure Weighted Rating (EWR)

Although the priority will always be to reduce the underestimation error, an additional complementary metric is proposed to address the overestimation problem.

$$EWR = \frac{\sum_{e,t}\left(\text{EXPOSURE}_{e,t} \times \text{RATING}_{e,t}\right)}{\sum_{e,t} \left(\text{EXPOSURE}\right)_{e,t}}, \forall (e,t) \in P,$$

where *e* represents an enterprise, *t* represents a month, and the pair *(e,t)* represents enterprise *e* observed in month *t*. This metric provides a sense of the overall conservatism of the model regarding the portfolio. For all existing enterprises in the portfolio, for all observed months, each enterprise's rating in each month is multiplied by its exposure in the same month, this product is summed across the portfolio and divided by the sum of all exposures across time

and companies. This measure does not provide any information on its own and should be used to complement the previous one.
For example, a model that predicts a rating of 9 for the entire portfolio will have EWDD equal to 0 for any *n* considered, whereas this model would have the highest EWR, indicating its pessimism.

One limitation that emerged during data preparation process was the non-random missingness of the outstanding loan amount, exposure. Specifically, the missingness within the group of enterprises that had defaulted at least once was less than 4%, whereas it was more than 85% for those that had never defaulted. This non-random absence of data could potentially introduce a significant bias into the EWR if it were calculated using imputed missing data based mainly on defaulted firms. To mitigate this problem, the volume of assets of the firms, ASSET_AMT, was used as a proxy measure, assuming that firms with larger asset volumes are likely to have higher exposures.
The EWDD, which is calculated using exposure data from defaulted firms, was accurately determined for those firms with recorded exposure information.

### 3.5.7. Model Comparison

In terms of the exposed metrics, the developed models, during the cross-validation process, were tuned taking into account the AUC. The best configuration of each type of model was used to estimate its performance on the unseen test dataset. This performance was assessed by their discriminative power (AUC). In addition, the models were compared according to the EWDD, for n equal to 3, 6 and 9 months, and the EWR.
In this study, all models were trained with the same features as the traditional model, which allowed the direct implementation of the proposed models, avoiding the collection and cleaning of new data.

## 4. Data

### 4.1. Data Description

The dataset used in this study was kindly provided by a Portuguese bank and includes a comprehensive historical record of the bank's SME portfolio over a five-year period, from

October 2018 to October 2023. In addition, the dataset includes the predicted ratings derived from the bank's operational model. This operational model is based on logistic regression, without regularisation techniques, and is complemented by a set of rules that can worsen the final rating.

The dataset prepared for training the models consists of an imbalanced panel data, containing a total of 931,109 observations across 34,904 unique enterprises. These enterprises are tracked over time, each operating independently without hierarchical interconnections. The dataset is characterized by its diversity in the number of months observed for each enterprise, contributing to its imbalanced nature. Each enterprise's record in the dataset includes crucial information such as its default status (0 representing “not in default” and 1 indicating “in default”), the date corresponding to this status, a credit rating predicted by the operational model, and the features employed by this model. Furthermore, the dataset incorporates the exposure value, indicative of the outstanding loan amount. See Table 3 for a complete list of these elements.

In contrast to conventional time series approaches, the models developed in credit risk assessment typically treat each observation as independent, bypassing the sequential aspects inherent in panel data.

Moreover, the dataset shows a significant class imbalance, a common attribute in credit risk analysis. This imbalance is visually represented in Figure 10, which shows the evolution of the SME loan portfolio over time. The graph includes a red line representing the total number of enterprises in the portfolio, highlighting an increasing trend. It also shows the number of enterprises that are not in default (shown in blue) and those that are in default (shown in orange). Over the period considered, these colour-coded lines show a gradual decline in the proportion of companies in default relative to the total portfolio.

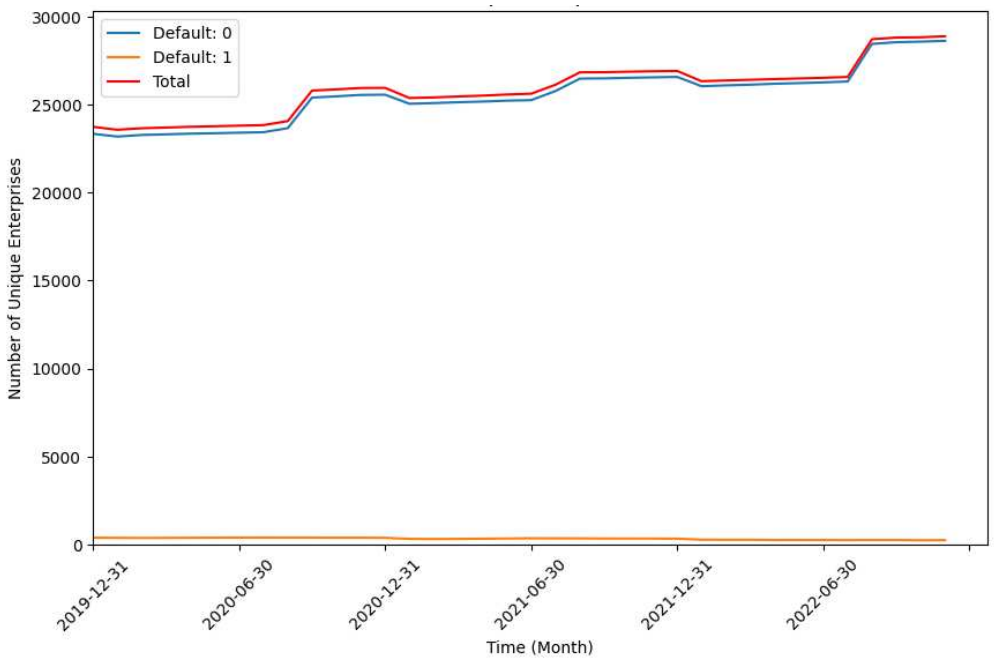


*Figure 10 - SME portfolio overview*

Looking in more detail at the trajectory of defaulted companies, Figure 11 shows a declining proportion over the analysed period. The data points start at a default rate of around 1.6% on December 2019 and show a declining pattern to almost 0.9% by November 2022. Despite minor fluctuations, the overall trend indicates a substantial reduction in defaults, which could reflect positively on the health and creditworthiness of the SME portfolio under scrutiny.

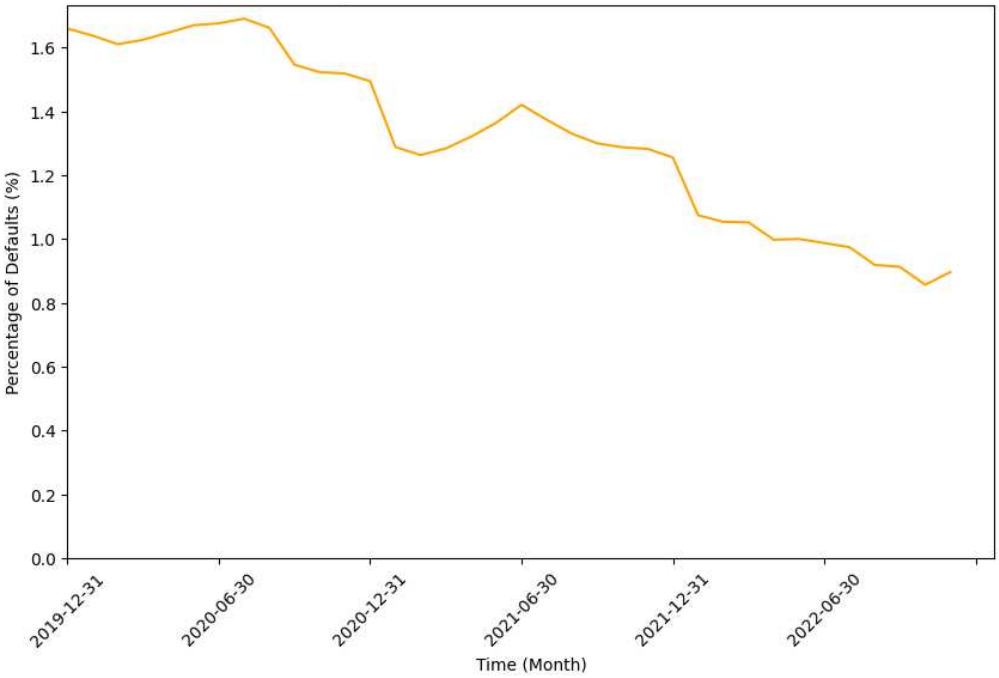


*Figure 11 - Proportion of defaulted Enterprises over time*

The primary objective of the predictive models is to forecast the *Target* variable, which is defined as a corporate default occurring within a 12-month horizon. When this specific variable is considered, the imbalance in the data set continues to follow the same trend Figure 12 .

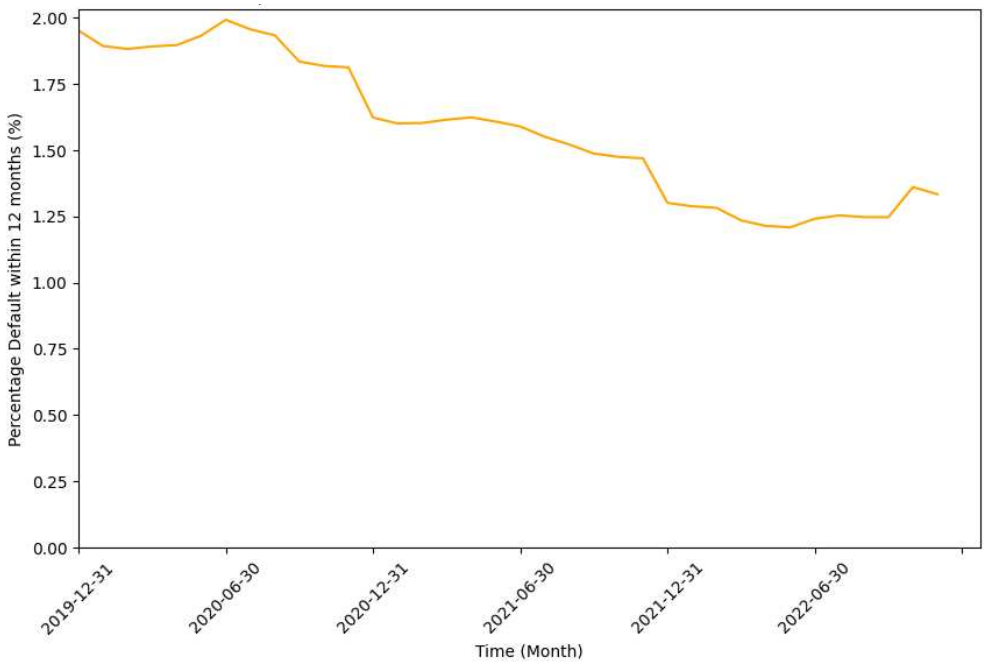


*Figure 12 - Proportion of Enterprises to default within 12 months Over Time*

The following violin plot  Figure 13 illustrates the evolution of the predicted ratings of the enterprises given by the operational model over the period under review.

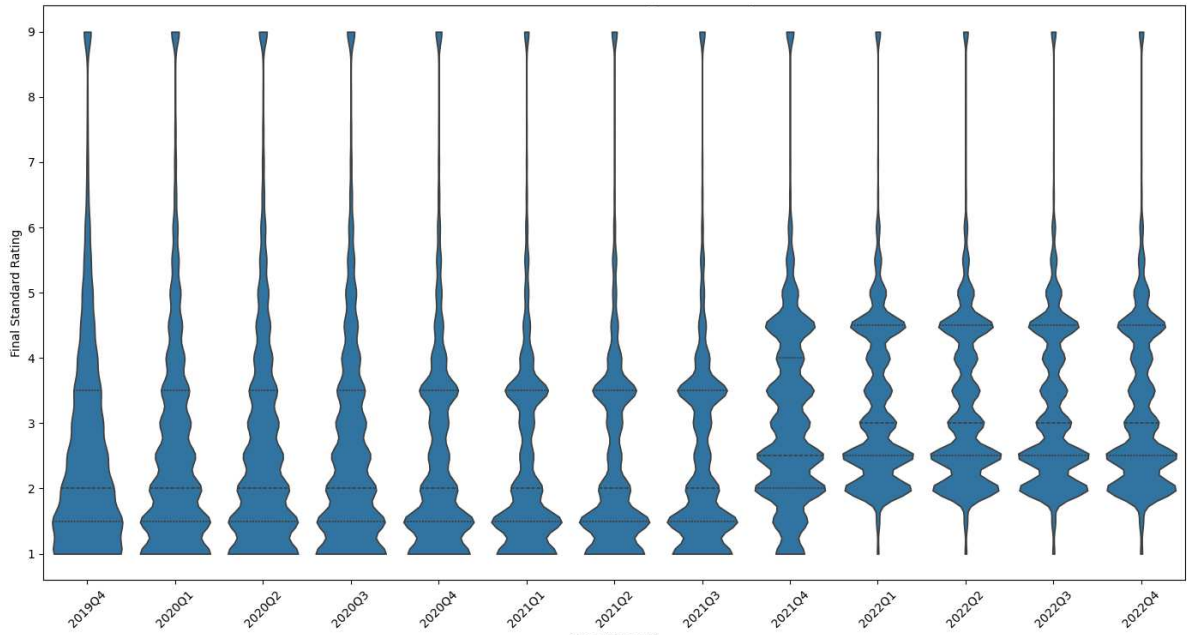


*Figure 13 - Violin Plot of Final Standard Rating Over Time by Quarter*

The financial indicators, the covariates, for the enterprises in this dataset show consistency over time, an expected observation given the dataset's focus on a specific market segment.

## 4.2. Data Preparation

### 4.2.1. Cleaning

Despite its structured format, the dataset presented several systematic challenges and issues that required preparation. The credit ratings of the enterprises in this portfolio are generally calculated on a monthly basis. However, the *CLASSIFICATION*_Q column, which records the questionnaire-based classification of each company, shows some irregularities. These inconsistencies are mainly due to the manual entry process carried out by the analysts and result in spontaneous recalculations of the ratings, which subsequently lead to additional entries for the enterprises concerned. In order to ensure consistency and accuracy in this study, a specific approach was adopted when dealing with multiple *CLASSIFICATION_Q* entries for the same enterprise on the same date, only the lowest value was retained. This method is deliberately

chosen to provide a conservative assessment of risk, in line with the precautionary principle typically applied in credit evaluation.

### 4.2.2. Missing data

The dataset presented a significant challenge with extensive missing data, particularly in the indicators, i.e. binary variables signalling the verification of certain conditions. These variables were either missing or active (represented by '1'), with no occurrences of '0'. To address this, all missing values in these columns were imputed with '0'. In addition, a threshold criterion for missingness was applied, any row exhibiting more than 70% missing features was excluded to maintain data quality.

In the refined dataset, a tailored approach was implemented to address remaining missing values. The dataset, which included numerical and categorical variables, required different treatment methods for these distinct types.

For numerical variables, missing values were interpolated using a linear method. This technique estimates missing values by identifying and leveraging a linear relationship between the known data points, ensuring a logical and statistically coherent imputation.

A different approach was taken for categorical variables. Missing values in these columns were replaced by the most common category within each group. This method was chosen as it preserves the underlying distribution of categorical data, thus maintaining the authenticity of the dataset.

Furthermore, the interpolation process was conducted within subsets of the dataset corresponding to the same enterprise. This grouping ensured that the interpolation of values was contextually relevant and maintained the integrity and coherence of the data.

The remaining missing values were filled in the final round of imputation, using the median of the column, in case of numerical, and with the category *'UNKOWN'*, in case of categorical features.

This comprehensive approach to dealing with missing data ensured that the dataset remained robust and representative for subsequent analyses, allowing more accurate and reliable findings to be drawn from the study.

### 4.2.3. Filtering

For several reasons, the datasets contained some firms with less than 12 months of observations. Given that the primary objective of the models is to estimate the probability of default over a 12-month period, the inclusion of these firms with incomplete observation histories was considered to potentially compromise the accuracy and reliability of the model. Consequently, to ensure the robustness and validity of the model's predictions, firms with less than 12 months of observations were excluded from the dataset.

### 4.2.4. Target Variable Creation

In the study, a pivotal step involved the creation of a *Target* variable to align with the model's primary objective of estimating the probability of default over a 12-month period for each enterprise. This process entailed an examination of each enterprise's default status across a rolling 12-month window, following each observation date.

The core of this procedure involved iterating through each record in the dataset. For every observation pertaining to an enterprise, a subsequent 12-month timeframe was inspected to detect any instances of default, this examination was confined to the same enterprise.

In case, within this 12-month window, at least one occurrence of default was observed, the *Target* variable for that particular observation date was marked as 1, indicating a predicted default within the ensuing year.

This approach to defining the Target variable was instrumental in aligning the dataset with the specific predictive objectives of the model, thereby providing a robust basis for subsequent modelling.

### 4.2.5. Final Filtering

In order to further refine the dataset for the modelling part, a crucial step was to exclude the last 11 months of observations. This decision was based on the inability to observe a full subsequent 12-month period within the time frame of the dataset. As a result, the future default status of these firms could not be determined, making the precise formulation of a target variable for these cases impractical. In addition, the first year of observations was found to be incomplete. While these initial observations were critical to the construction of the target variable, they were deemed unsuitable for model estimation due to their partial nature. These exclusions were essential to maintaining the integrity and reliability of the modelling process,

ensuring that all data used for training and analysis had a complete and informative set of characteristics, including a robustly derived Target variable.

## 5. Results

In this study, the optimal configurations of each class of predictive models were selected to evaluate their performance on a test dataset representing 30% of all enterprises. To ensure an impartial evaluation, this test dataset was strictly segregated during the training process, ensuring an unbiased assessment of the predictive capabilities of the models. The test data consists of 279,905 observations representing 10,472 SMEs, of which 282 have recorded defaults.

### 5.1. Assessing Performance using Commonly Used Metrics

In alignment with the established practices in credit risk modelling, as discussed in section 3.5. Metrics, the AUC was employed as the principal metric to evaluate the discriminative power of each model. Subsequent Figure 14 delineates the ROC curves and their corresponding AUC values, for each model. All the models have demonstrated substantial predictive power with Random Forest (RF) and XGBoost (XGB) leading in terms of AUC achieving, 0.98 and 0.97 respectively. These findings are consistent with prior research (Bhatore et al., 2020; Noriega et al., 2023; Shi et al., 2022; Siddharth Bhatore et al., 2020). The KNN, the simple logistic regression (LR) and the hybrid, decision tree and logistic regression (DT+LR), indicate the same AUC, 0.96, i.e., it was not observed a substantial increase of AUC between the LR and the hybrid DT+LR models. The SVC model was outperformed presenting the lowest AUC of 0.88. Notably, the ambitious hybrid model combining the neural network with the logistic regression (NN+LR), contrary to expectations, yielded a relatively modest AUC of 0.94.

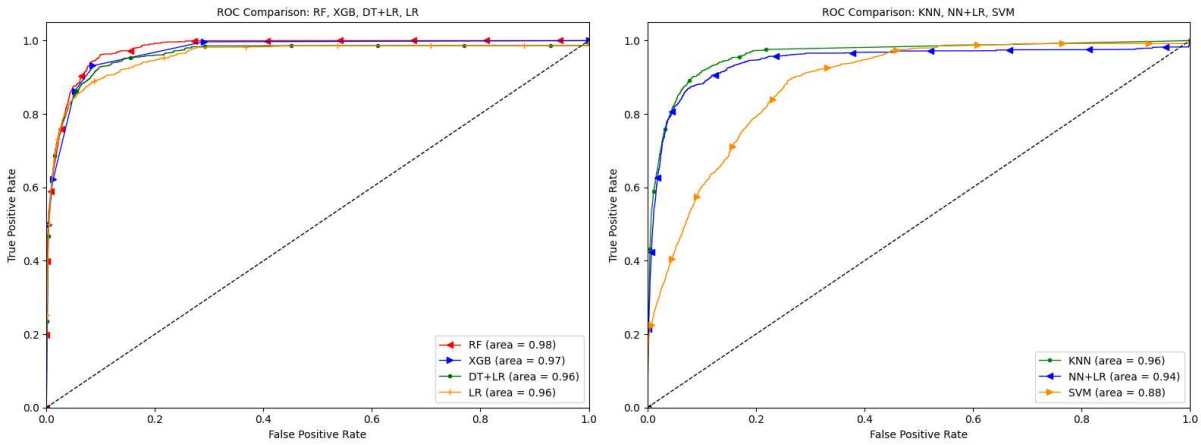


*Figure 14 - ROC Curves of the Developed Models*

In evaluating the performance of the XGB, DT+LR and RF models using confusion matrixes Table 1, it is observed that the XGB model has a robust ability to identify true negative instances (252153) and maintains a respectable rate of true positive identifications (4081). However, it also presents a marked number of false positives (23345), which requires careful consideration of the implications associated with false positive classifications. The DT+LR model has an increased propensity for false positive predictions (45979), which is substantially higher than the XGB model, with a slight increase in true positives (4209). This suggests a trade-off between sensitivity and specificity. In contrast, the Random Forest model achieves a closer balance with 247174 true negative and 4240 true positive predictions but does not substantially reduce the number of false positives (28324) compared to the XGB model. The selection of the most appropriate model for use should be based on the relative costs of false negative and false positive predictions and the overall objective of the credit risk assessment process within the SME segment.

*Table 1 - Confusion Matrixes*

**XGBoost Confusion Matrix**

| | Predicted Negative | Predicted Positive |
|---|---|---|
| Actual Negative | 252153 | 23345 |
| Actual Positive | 326 | 4081 |

**Decision Tree Logistic Regression Hybrid Confusion Matrix**

| | Predicted Negative | Predicted Positive |
|---|---|---|
| Actual Negative | 229519 | 45979 |
| Actual Positive | 198 | 4209 |

**Random Forest Confusion Matrix**

| | Predicted Negative | Predicted Positive |
|---|---|---|
| Actual Negative | 247174 | 28324 |
| Actual Positive | 167 | 4240 |

### 5.2. Assessing Performance using Exposure Weighted Metrics

Following the evaluation of the commonly used metrics, the predicted probabilities were then mapped onto the master scale, using the same criteria as the operational model, thereby enabling the calculation of the Exposure Weighted Ratings (EWR) and the Exposure Weighted Distance to Default (EWDD) for 3, 6, and 9 months. The *FINAL_RATING_STANDARD* serves as a benchmark, representing the ratings predicted by the operational model. See Table 4 in 8. Appendix to observe the calculated metrics for each model.

**Error! Reference source not found.** illustrates the observed metrics of each model developed. N otably, XGB, RF and the hybrid DT+LR model demonstrate superior performance to the operational model in terms of EWDD across all examined time frames and lead to a lower EWR. The figure reveals that the XGB model, despite not having the highest AUC, demonstrates the lowest EWDD across all considered periods. This finding indicates that the XGB model is particularly adept at predicting higher ratings for enterprises shortly before default, when weighted by exposure.

Furthermore, the hybrid DT+LR model, which has the same AUC of 0.96 as the simple LR model, outperforms it and demonstrates comparable performance to the RF model as per the metrics proposed.

The highest EWR observed corresponds to the operational model and considering its EWDD results one can infer that the operational model compensates its relatively low predictive power assigning overall high ratings.

Although the simple LR, KNN, SVM, and hybrid NN+LR models exhibited a considerably lower EWR than the operational model, they were unable to outperform it in terms of EWDD. The unexpected results of the hybrid NN+LR model are further discussed in section 6. Conclusion and Limitations.

Additionally, it should be noted that of the three models surpassing the traditional model in all proposed metrics, the DT+LR offers an interpretability level comparable to that of the simple logistic regression.

For an alternative visual comparison of the observed metrics for the operational, XGB, RF and DT+LR models see 8. Appendix Figure 16.

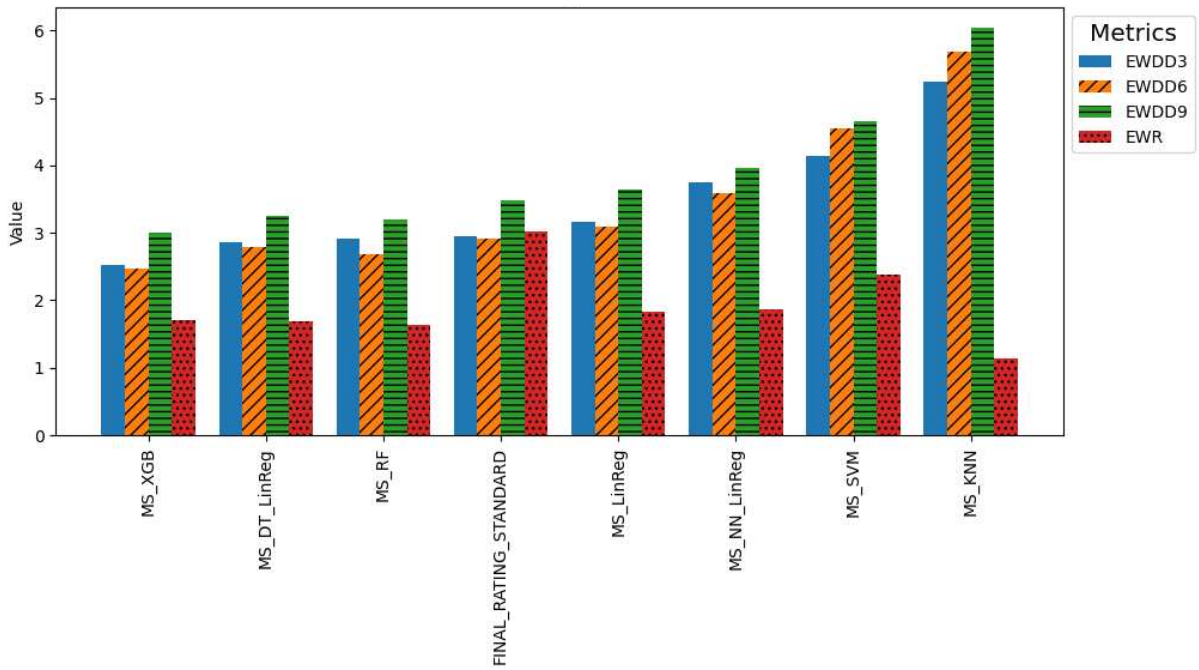


*Figure 15 - Observed metrics for each model*

Table 2 illustrates the percentage change in each metric relative to the *FINAL_RATING_STANDARD,* operational model. The XGB achieves more than a 13.4%

reduction in EWDD for all periods considered. In EWDD3, the hybrid DT+LR model outperforms the RF model with a reduction of 3.3% and 1.3% respectively. However, in EWDD6, the RF model achieves a better result with a reduction of 7.9% compared to the DT+LR model's 4.2%. Similarly, in EWDD9, the RF model outperforms the DT+LR model with a reduction of 8.0% compared to 6.6%.

This indicates that these three models would be able to perform better in estimating the capital requirements for those firms that defaulted requiring a higher capital buffer for them. Moreover, XGB, DT+LR and the RF models are more then 40% less conservative when assessing the overall portfolio, according to EWR, which leads to a substantial reduction in overall capital requirements.

*Table 2 - Percentage change in each metric relative to the operational model*

| MODEL | EWDD3 | EWDD6 | EWDD9 | EWR |
|---|---|---|---|---|
| MS_XGB | -14.6% | -15.4% | -13.4% | -43.2% |
| MS_DT_LR | -3.3% | -4.2% | -6.6% | -43.9% |
| MS_RF | -1.3% | -7.9% | -8.0% | -46.0% |
| FINAL_RATING_STANDARD | 0.0% | 0.0% | 0.0% | 0.0% |
| MS_LR | 7.5% | 6.0% | 4.7% | -39.3% |
| MS_NN_LR | 27.0% | 22.8% | 13.9% | -38.0% |
| MS_SVM | 40.5% | 55.7% | 33.8% | -21.0% |
| MS_KNN | 77.5% | 94.7% | 73.6% | -62.2% |

In summary, the results suggest that XGB, DT+LR and the RF models were able to better discriminate between enterprises that were close to default, requiring an even higher capital buffer for them than the operational model, while at the same time requiring less capital retention for enterprises that had not defaulted.

## 6. Conclusion and Limitations

This study evaluated SME credit rating models using AUC, EWR and EWDD and found that exposure weighted metrics lead to more informed model selection, highlighting the predictive strengths and weaknesses of each model.

The results highlight that while the AUC is a robust and informative metric for assessing the predictive power, the introduction of exposure weighted metrics reveals a nuanced view of

model performance pointing out the benefits of using ML techniques. These metrics outline the fact that not all incorrect predictions carry the same business impact leading to a more accurate model comparison. The proposed metrics were able to outline, that in this case, in terms of indirect estimation of capital requirements, XGB, DT+LR and the RF models can lead to a more accurate assessment of credit risk, suggesting an increase in capital retention for enterprises close to default and a substantial decrease in capital retention for those not close to default. The overall reduction of the capital buffer, up to 46% according to EWR, could potentially lead to additional value generation by improving the access to finance of SMEs, in line with the regulators' objectives (EBA, 2023; EBA, 2016), and increasing the financial institution's profitability. None of the other models have shown a reduction in EWDD across any of the time frames considered. As the primary objective of a financial institution's credit risk assessment is to maintain its economic stability by avoiding the underestimation of borrower default probabilities, none of them, as they are, can be considered as candidates to substitute the operational model.

The study confirms that ensemble methods outperform in credit risk assessment, excelling in both predictive power and exposure weighted metrics. This aligns with previous findings, (Bhatore et al., 2020; Bussmann et al., 2021a; de Lange et al., 2022) highlighting their efficiency in accurately evaluating credit risk. A standout finding is the efficacy of a hybrid decision tree and logistic regression model, offering a pragmatic balance between analytical depth and relieve of interpretation complexity in credit risk analysis. The hybrid DT+LR model demonstrated impressive results, achieving performance levels comparable to ensemble methods while maintaining a relatively low level of interpretability complexity. This balance between effectiveness and usability highlights its potential value in practical credit risk assessment applications. Although the XGB model exhibits superior performance, the financial institution must weigh the benefits of this marginal improvement against the increased complexity in interpretability. This consideration is fundamental in deciding whether the enhanced performance justifies the potential challenges in model understanding and explanation.

In summary, this study has evaluated the impact of using different ML techniques in credit risk estimation, providing two comprehensive metrics that indicate the benefits of using these techniques. Additionally, this study addressed the complexity of interpretability and

explainability of some ML models by developing a hybrid model with relatively low complexity and ambitious results.

Despite implementing strategies to mitigate overfitting and data imbalance, the hybrid model combining neural network and logistic regression was outperformed. Potential factors for this outcome could include suboptimal hyperparameter tuning, inappropriate model architecture, or ineffective feature extraction from the NN. Time constraints limited an exhaustive investigation of these challenges, highlighting the complexities involved in optimizing such hybrid models.

The study encountered several limitations, including data quality issues that necessitated assumptions, a limited set of features constrained to those employed by the operational model, and a relatively short span of observational data, only five years, with just three suitable for modelling. The non-random missingness of the exposure variable impeded its direct use in calculating the Exposure Weighted Rating (EWR), necessitating the adoption of an alternative financial indicator as a proxy. A further limitation of the study is that the period covered includes the COVID-19 pandemic and the subsequent post-pandemic recovery period. This is concerning because the economic and financial conditions during this period were highly atypical, with substantial disruption to business operations, changes in consumer behaviour and exceptional government intervention. These factors can create anomalies in the data that may not be representative of normal market conditions, potentially distorting the model's predictions and affecting its generalisability to other periods.

The limitations of this study, particularly regarding data quality and volume, lead to restrictions in the results. This suggests the need for further research with more comprehensive, larger, and higher quality datasets to validate whether the conclusions drawn, especially regarding the performance of the DT+LR model, are consistent with the outcomes observed in this study. Furthermore, following the approach in (de Lange et al., 2022), the proposed hybrid model should be evaluated and compared with other models based on Loss Given Default. Additionally, it's important to incorporate relevant stress tests commonly used in financial institutions to estimate the stability and robustness of this model under various economic scenarios. These additional analyses would provide a more comprehensive understanding of the applicability and reliability of this model in real-world financial contexts.

## 7. References


Abdulsaleh, A. M., & Worthington, A. C. (2013). Small and Medium-Sized Enterprises Financing: A Review of Literature. *International Journal of Business and Management*, *8*(14). https://doi.org/10.5539/ijbm.v8n14p36

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, *23*(4), 589. https://doi.org/10.2307/2978933

Altman, E. I. (2018). A fifty-year retrospective on credit risk models, the Altman Z -score family of models and their applications to financial markets and managerial strategies. *The Journal of Credit Risk*, *14*(4), 1–34. https://doi.org/10.21314/JCR.2018.243

Altman, E. I., & Sabato, G. (2005). *Modeling Credit Risk for SMEs: Evidence from the US Market*. www.sba.gov.

Baehrens, D., Harmeling, S., Kawanabe, M., Hansen KHANSEN, K., & Edward Rasmussen, C. (2010). How to Explain Individual Classification Decisions Timon Schroeter * Klaus-Robert M ¨ uller. In *Journal of Machine Learning Research* (Vol. 11).

Bańkowska, K., Ferrando, A., & Garcia, J. A. (2020). *Access to finance for small and medium-sized enterprises since the financial crisis: evidence from survey data*. https://www.ecb.europa.eu/pub/economic-bulletin/articles/2020/html/ecb.ebart202004_02~80dcc6a564.en.html

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bhatore, S., Mohan, L., & Reddy, Y. R. (2020). Machine learning techniques for credit risk evaluation: a systematic literature review. *Journal of Banking and Financial Technology*, *4*(1), 111–138. https://doi.org/10.1007/s42786-020-00020-3

Breiman, L. (2001). Random Forest. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification And Regression Trees*. Routledge. https://doi.org/10.1201/9781315139470

Burkart, N., & Huber, M. F. (2021). A Survey on the Explainability of Supervised Machine Learning. *Journal of Artificial Intelligence Research*, *70*, 245–317. https://doi.org/10.1613/jair.1.12228

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021a). Explainable Machine Learning in Credit Risk Management. *Computational Economics*, *57*(1), 203–216. https://doi.org/10.1007/s10614-020-10042-0

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021b). Explainable Machine Learning in Credit Risk Management. *Computational Economics*, *57*(1), 203–216. https://doi.org/10.1007/s10614-020-10042-0

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. https://doi.org/10.1145/2939672.2939785

Clement, T., Kemmerzell, N., Abdelaal, M., & Amberg, M. (2023). XAIR: A Systematic Metareview of Explainable AI (XAI) Aligned to the Software Development Process. *Machine Learning and Knowledge Extraction*, *5*(1), 78–108. https://doi.org/10.3390/make5010006

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random Forests. *Ensemble Machine Learning*, 157–175. https://doi.org/10.1007/978-1-4419-9326-7_5

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, *2*(4), 303–314. https://doi.org/10.1007/BF02551274

de Lange, P. E., Melsom, B., Vennerød, C. B., & Westgaard, S. (2022). Explainable AI for Credit Assessment in Banks. *Journal of Risk and Financial Management*, *15*(12), 556. https://doi.org/10.3390/jrfm15120556

Delgado-Panadero, Á., Hernández-Lorca, B., García-Ordás, M. T., & Benítez-Andrades, J. A. (2022). Implementing local-explainability in Gradient Boosting Trees: Feature Contribution. *Information Sciences*, *589*, 199–212. https://doi.org/10.1016/j.ins.2021.12.111

EBA. (2016). *EBA REPORT ON SMES AND SME SUPPORTING FACTOR*.

EBA. (2020). *EBA REPORT ON BIG DATA AND ADVANCED ANALYTICS EBA REPORT ON BIG DATA AND ADVANCED ANALYTICS 2*.

EBA. (2021). *REPORT ON MACHINE LEARNING IN CREDIT RISK EBA DISCUSSION PAPER ON MACHINE LEARNING FOR IRB MODELS*.

EBA. (2023). *MACHINE LEARNING FOR IRB MODELS FOLLOW-UP REPORT FROM THE CONSULTATION ON THE DISCUSSION PAPER ON MACHINE LEARNING FOR IRB MODELS*.

ECB. (2023). *ECB guide to internal models*.

Eugene Mazur. (2012). *"Green Transformation of Small Businesses: Achieving and Going Beyond Environmental Requirements", OECD Environment Working Papers, No. 47, OECD Publishing*.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, *4*(37). https://doi.org/10.1126/scirobotics.aay7120

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023a). *Classification* (pp. 135–199). https://doi.org/10.1007/978-3-031-38747-0_4

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023b). *Linear Model Selection and Regularization* (pp. 229–288). https://doi.org/10.1007/978-3-031-38747-0_6

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023c). *Statistical Learning* (pp. 15–67). https://doi.org/10.1007/978-3-031-38747-0_2

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023d). *Support Vector Machines* (pp. 367–398). https://doi.org/10.1007/978-3-031-38747-0_9

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023e). *Tree-Based Methods* (pp. 331–366). https://doi.org/10.1007/978-3-031-38747-0_8

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023f). *Tree-Based Methods* (pp. 331–366). https://doi.org/10.1007/978-3-031-38747-0_8

Kraemer-Eis, Lang, H. ;, & Frank. (2012). *The importance of leasing for SME finance*. http://hdl.handle.net/10419/176645

Lloyd Shapley. (1953). *A Value for N-Person Games, Contributions to the Theory of Games*.

Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2019). *Explainable AI for Trees: From Local Explanations to Global Understanding*.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Noriega, J. P., Rivera, L. A., & Herrera, J. A. (2023). Machine Learning for Credit Risk Prediction: A Systematic Literature Review. *Data*, *8*(11), 169. https://doi.org/10.3390/data8110169

Rao, P., Kumar, S., Chavan, M., & Lim, W. M. (2023). A systematic literature review on SME financing: Trends and future directions. *Journal of Small Business Management*, *61*(3), 1247–1277. https://doi.org/10.1080/00472778.2021.1955123

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016a). *Model-Agnostic Interpretability of Machine Learning*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016b). "Why Should I Trust You?" *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Roy, P. K., & Shaw, K. (2021). A multicriteria credit scoring model for SMEs using hybrid BWM and TOPSIS. *Financial Innovation*, *7*(1), 77. https://doi.org/10.1186/s40854-021-00295-5

Rozemberczki, B., Watson, L., Bayer, P., Yang, H.-T., Kiss, O., Nilsson, S., & Sarkar, R. (2022). *The Shapley Value in Machine Learning*.

Sagi, O., & Rokach, L. (2021). Approximating XGBoost with an interpretable decision tree. *Information Sciences*, *572*, 522–542. https://doi.org/10.1016/j.ins.2021.05.055

Salih, A., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Menegaz, G., & Lekadir, K. (2023). *Commentary on explainable artificial intelligence methods: SHAP and LIME*.

Segal, M., & Xiao, Y. (2011). Multivariate random forests. *WIREs Data Mining and Knowledge Discovery*, *1*(1), 80–87. https://doi.org/10.1002/widm.12

Shi, S., Tse, R., Luo, W., D'Addona, S., & Pau, G. (2022). Machine learning-driven credit risk: a systemic review. *Neural Computing and Applications*, *34*(17), 14327–14339. https://doi.org/10.1007/s00521-022-07472-2

Siddharth Bhatore, Lalit Mohan, & Y. Raghu Reddy. (2020). *Machine learning techniques for credit risk evaluation: a systematic literature review*.

Silva Filho, T., Song, H., Perello-Nieto, M., Santos-Rodriguez, R., Kull, M., & Flach, P. (2023). Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Machine Learning*, *112*(9), 3211–3260. https://doi.org/10.1007/s10994-023-06336-7

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: the Achilles heel of predictive analytics. *BMC Medicine*, *17*(1), 230. https://doi.org/10.1186/s12916-019-1466-7

Zhang, Z., Zhang, T., & Li, J. (2023). *Unbiased Gradient Boosting Decision Tree with Unbiased Feature Importance*. http://arxiv.org/abs/2305.10696

Zhao, X., & Nie, X. (2021). Splitting Choice and Computational Complexity Analysis of Decision Trees. *Entropy*, *23*(10), 1241. https://doi.org/10.3390/e23101241

## 8. Appendix

*Table 3 - Data Dictionary*

| Variable | Label |
|---|---|
| IDENTIFIER | Anonymous Company Identifier |
| DEFAULT_FLG | Default Status |
| FINAL_RATING_STANDARD | Final Standard Rating |
| LOAD_DTTM | Analysis Date |
| ASSET_AMT | Asset Volume |
| AUTONOMIA_FINAN_CORRIGIDA_PCT | Corrected Financial Autonomy Percentage |
| AVG_CRED_VENC_VS_RESP_BANCA_12M | Average Past Due Credit / Bank Liabilities (12m Average) |
| AVG_CRED_VENC_VS_RESP_CGA_12M | Average Past Due Credit / CGA Liabilities (12m Average) |
| CLASSIFICATION_Q | Analyst Classification |
| EXPOSURE | Exposure Value |
| FLAG_DESCOBERTOS_N_AUT_12M | Indicator of Unauthorized Overdraft in the Last 12m |
| FLAG_ESFL | Non-Profit Institutions Indicator |
| FLAG_SGPS | Holding Companies Indicator |
| FLAG_STARTUP | Startup Indicator |
| INDICADOR_CRED_VENC_BANCA_12M | Indicator of Past Due Credit in the Last 12 Months at the Bank |
| INDICADOR_CRED_VENC_GCA_12M | Indicator of Past Due Credit in the Last 12 Months at CGA |
| JUROS_GASTOS_SUPORTADOS_VS_ATIVO | Interest and Similar Charges Supported / Total Assets |
| MARGEM_EXPLO_JUROS_SUPORT | Gross Operating Margin / Supported Interest and Similar Charges |
| MEIOS_FINAN_PASSIVO_CORRENTE | (Liquid Financial Resources + Current Inventories and Biological Assets) / Current Liabilities |
| NET_DEBT_EBITDA | Net Debt / EBITDA Ratio |
| PD_CALIBRADA_FINAL | Final Calibrated Probability of Default |
| PRAZO_MEDIO_RECEBIMENTO | Average Receivables Period (days) |
| RESPONS_BANCA_VS_EBITDA | Bank Liabilities / EBITDA |
| RESPONS_BANCA_VS_RENDIMENTO | Bank Liabilities / Total Income |
| RESPONS_GCA_VS_RENDIMENTO | CGA Liabilities / Total Income |
| RESULTADO_OPER_VS_JUROS_SUPOR | Operating Result / Supported Interest and Similar Charges |
| RESUL_ANTES_IMP_VS_RESP_BANCA | Pre-Tax Result / Bank Liabilities |
| SECTION_ECONOMIC_ACTIVITY_CD | Economic Activity Section Code |
| TOTAL_FINAN_OBT_VS_TOTAL_PASSIVO | Total Financing Obtained / Total Liabilities |
| TURNOVER_AMT | Business Volume |
| UNEMPLOYMENT_RT | Unemployment Rate |
| VOLUME_NEG_VS_ACT_CORR | Business Volume / Current Assets |
| X_UTR_CD_12M | Risk User at BdP at Least Once in the Last 12m |

*Table 4 - Measuring the business impact using EWDD and EWR*

| MODEL | EWDD3 | EWDD6 | EWDD9 | EWR |
|---|---|---|---|---|
| MS_XGB | 2.518697 | 2.467231 | 3.008609 | 1.715489 |
| MS_DT_LR | 2.852526 | 2.795557 | 3.244707 | 1.694191 |
| MS_RF | 2.910339 | 2.688763 | 3.198908 | 1.631414 |
| FINAL_RATING_STANDARD | 2.948870 | 2.917899 | 3.475747 | 3.021267 |
| MS_LR | 3.169504 | 3.093784 | 3.640621 | 1.835005 |
| MS_NN_LR | 3.745882 | 3.582538 | 3.957678 | 1.873206 |
| MS_SVM | 4.144051 | 4.543564 | 4.650761 | 2.387380 |
| MS_KNN | 5.233680 | 5.681807 | 6.033033 | 1.143034 |

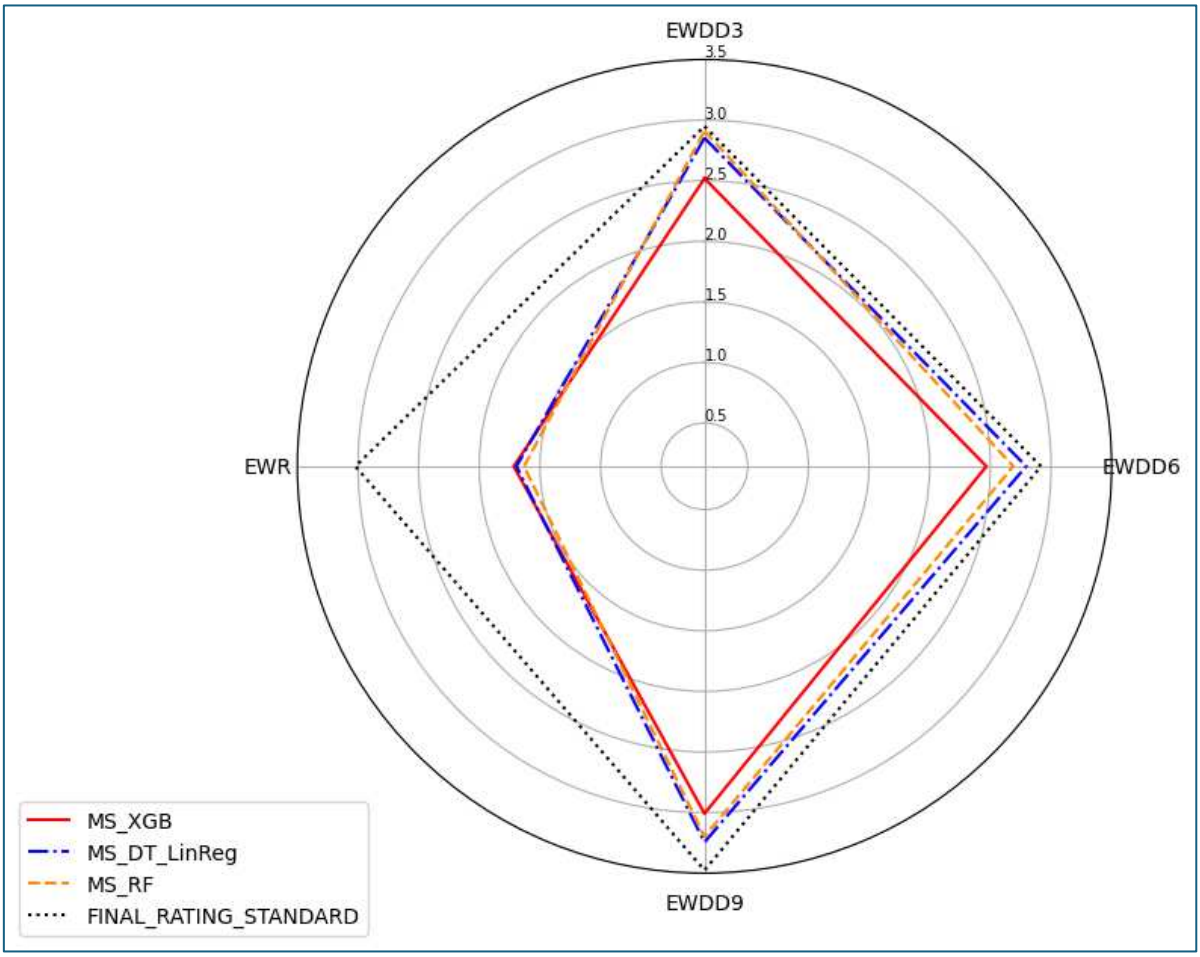


*Figure 16 - Radar plot illustrating the performance of the models in exposure weighted metrics*