ETHICAL TRANSPARENCY IN BUSINESS FAILURE PREDICTION: UNCOVERING THE BLACK BOX OF XGBOOST ALGORITHM

Authors: Mariano Romero Martínez, José Pozuelo Campillo & Pedro Carmona Ibáñez (University of Valencia)

ABSTRACT

This paper introduces an innovative machine learning model, built upon the robust and interpretable Extreme Gradient Boosting (XGBoost technique), to predict business failure. The study uses the VADIS financial strength indicator, from ORBIS, in a sample of 3,806 Spanish firms. The empirical analysis demonstrates that higher solvency, profitability, and reduced indebtedness correlate with a lower propensity for business failure. The model offers a powerful tool for early detection and intervention in financial distress, helping prevent insolvency.

This study extends its focus to the ethical dimensions inherent in the application of machine learning algorithms. It attempts to demystify the proverbial "black box" nature of such techniques, shedding light on their ethical foundations. By demonstrating the transparency and interpretability of the XGBoost model, we address the ethical concerns often associated with the "black box" nature of machine learning algorithms.

KEYWORDS: Business failure, prediction models, machine learning, XGBoost, ethics of algorithms.

Financial support from MCIU/AEI/10.13039/501100011033/FEDER/UE (Project PID2023-153128NB-I00) is gratefully acknowledged.

1. INTRODUCTION

Economic crises, characterized by a surge in business failures, have historically attracted heightened research attention, prompting investigations into their causative factors, consequences, and methods of detection (Mousavi et. al., 2015). Traditionally, the methodological arsenal in this domain relied heavily on statistical techniques rooted in the established data modeling. These methods included discriminant analysis and logistic regression, which were deemed as classic tools for addressing these challenges.

However, in recent years, there has been a notable expansion of the methodological landscape within this field. This transformation has been driven by the infusion of cutting-edge technologies, particularly artificial intelligence (AI) and machine learning (ML), into the field of business failure prediction. The integration of AI and ML techniques has facilitated a more profound understanding of the underlying processes and, in many instances, has exceeded the predictive capabilities and applicative ease of conventional "classic" models, thus entering in the ambit of algorithms.

The arrival of AI and ML has enriched the methodological repertoire in business failure analysis, offering a diverse array of tools and approaches to explore this complex phenomenon. These techniques, characterized by their ability to learn from data patterns and adapt, have empowered researchers to explore deeper into the intricacies of business failure dynamics. By joining the power of AI and ML, scholars have gained access to predictive models that not only rival but often outperform traditional methods, providing more accurate and actionable insights into the likelihood of business failure.

These modern techniques have demonstrated a remarkable capacity to handle vast datasets and capture intricate relationships, thereby enhancing their usability in real-world scenarios. As a result, researchers and practitioners in the field of economic crises and business failure now have a wealth of advanced tools at their disposal to better comprehend, anticipate, and respond to the multifaceted challenges presented by economic downturns.

These advanced methodologies offer not only enhanced predictive capabilities but also greater versatility and adaptability, contributing to a deeper and more comprehensive understanding of the intricacies associated with business failure during periods of

economic turmoil. This evolution underscores the transformative potential of AI and ML in advancing our knowledge and capacity to address critical economic challenges.

Generally speaking, the analytical process of Machine ML involves algorithms that absorb information from the huge volume of data—both structured and unstructured—with which businesses are flooded on a daily basis, in order to extract information and predict complex phenomena (Oyewo et al., 2020; Rapanyane & Sethole, 2020). By including techniques of a dynamic nature that can continually run in the background, they provide company managers with relevant, real-time information that facilitates the decision-making process (Syam & Sharma, 2018). Notable related studies include that by Xia et al. (2017), who propose a credit scoring model that outperforms previous credit scoring systems in terms of reliability and accuracy, and the study of software companies by Santhanam et al. (2017), who conclude that the application of the Extreme Gradient Boosting (XGBoost) algorithm provides the most appropriate model to analyze the aspects that influence customer retention. Furthermore, according to Climent et al. (2019), this type of technique works better than classic linear regressions and other statistical techniques when it comes to issues related to business insolvency.

One of the most innovative ML applications is the aforementioned XGBoost, a supervised learning algorithm aimed at predicting a target variable with maximum accuracy by combining the estimations of a set of simpler and weaker models. The fact that this algorithm enables highly accurate, fast models has attracted the interest of researchers from numerous fields seeking to incorporate it into their studies. Indeed, for several years now it has been successfully used as a prediction technique in research areas such as medicine, for the diagnosis of various pathologies (Merlet et al., 2001; Dettling and Bühlmann, 2003; Rodríguez Jiménez, 2005; Sanz et al., 2015). However, it has only very recently started to be applied in economics, and in the specific field of business failure prediction there have been markedly few contributions to the literature.

One of the studies to use XGBoost in the field of business failure research is that by Climent et al. (2019), who use the algorithm to identify the financial indicators that could predict bankruptcy in the financial institutions of the eurozone. Another study similar in methodological terms but focuses on US financial institutions is the one carried out by

Carmona et al. (2019), who propose a battery of ideal indicators to predict bankruptcy in this type of company. These studies highlight the superiority of these techniques over the classic ones when it comes to predicting and explaining situations of financial difficulty in the analyzed companies.

A recently published paper by Carmona et al. (2022) proposes a novel way to discover the contents of the “black box” associated with XGBoost and shows how using this technique makes it possible to fit a highly accurate, easily interpretable ML model.

The fact is that the remarkable rise in the use of algorithms raises ethical implications in a society ever more governed by algorithms. For example, when someone is planning on booking a trip, artificial intelligence software evaluates the offers on the basis of the consumer's situation; when someone is planning to take out insurance, an algorithm compares prices; when a megastore recommends you buy a certain product, it is really an algorithm doing it; and when a company wants to take out a loan, artificial intelligence software performs a study of its options and offers certain alternatives depending on its financial guarantees. As such, it is very important to know and to understand who manages the information—a key aspect for individual and collective freedoms. Monasterio (2017), discusses the ethical implications of a society increasingly governed by algorithms. The author defines an algorithm as a software code that processes a limited set of instructions and highlights three key properties of algorithms: universality, opacity, and impact on people's lives. The author argues that algorithms are not neutral, objective, or pre-analytical, but are shaped by the ideas, practices, instruments, and contexts of their creators. The paper then presents several examples of how algorithms can discriminate against individuals based on their social, economic, and racial backgrounds. The author concludes by calling for the development of ethical guidelines and regulations for the use of algorithms to ensure that they are used in a fair and just manner.

Therefore, the increasing use of explainable AI (XAI) and ML models in financial decision-making processes has raised concerns about the lack of transparency and interpretability of these models. This lack of transparency, previously referred to as the ‘black box’ problem (Vega & Aznarte, 2020), can lead to a lack of trust and accountability, especially in high-stakes domains such as finance (Gilpin et al., 2018).

In response to these concerns, there has been a growing interest in developing XAI methods that can provide insights into how AI models make decisions (Carvalho et al., 2019).

The ethical implications of AI in finance are significant. The use of AI models in credit scoring, loan approvals, and investment decisions can have a profound impact on individuals and businesses. If these models are not transparent and interpretable, it can be difficult to ensure that they are fair, unbiased, and non-discriminatory (Doshi-Velez & Kim, 2017). Furthermore, the lack of transparency can make it difficult to hold individuals or organizations accountable for the decisions made by AI models.

The increasing prevalence of artificial intelligence and machine learning in finance has motivated a surge of interest in the interpretability of these models. Researchers and practitioners alike are recognizing the critical need to understand how these complex algorithms arrive at their conclusions, not only to ensure transparency and accountability but also to address potential biases and ethical concerns

Hence, this paper tries to reveal the inner workings of the "black box" in the domain of financial analysis. Our primary objective is to shed light on the specific ratios and indicators within this enigmatic box that serve as robust predictors of future financial distress. In doing so, we aim to equip company managers with a potent and practical tool that can help navigate their organizations away from the precipice of financial difficulty. This is a crucial departure from conventional approaches, as it underscores the objectivity of our methodology, reinforcing the credibility and reliability of our results. Our study explores new areas by connecting financial analysis with ethical objectivity

Our comprehensive review of the existing body of knowledge in this field has yielded an interesting observation. While previous studies have undertaken the field of financial analysis, few have dared to question the ethical foundations of the black box. While we acknowledge the valuable contributions made by studies such as those by Carmona et al. (2022) and Climent et al. (2019). Unlike them which focus on identifying the most relevant ratios, our study tries to go further by relating ML interpretability and the ethics of algorithms. We aspire to uncover the hidden mechanics of the black box

to attain the ethical objectivity embedded within its contents. That is, we seek to enhance the transparency and interpretability of the models we propose.

In our study, we employ a technique known as break-down plots. This approach allows us to dissect the probability of business failure assigned by the XGBoost model to each company in our dataset. By examining the contribution of each explanatory variable (feature) to the final prediction, we gain a thorough understanding of the factors driving the model's assessment.

In essence, we break down the probability of failure into its constituent parts, revealing the influence of variables such as solvency, profitability, and indebtedness. This not only enhances the transparency of our model but also empowers stakeholders with actionable insights. Understanding the specific factors that contribute to a company's risk profile enables decision-makers to make more informed choices and potentially intervene to mitigate the risk of financial distress

Furthermore, breaking down the probability of failure for each observation is crucial from an ethical standpoint. It ensures fairness and avoids discriminatory practices by revealing the specific factors influencing the model's decision. This transparency allows for scrutiny and challenges potential biases in the data or the model itself. It also promotes accountability, as stakeholders can trace the reasoning behind each prediction, fostering a responsible and ethical use of AI in financial assessments.

This study not only contributes to the growing body of research on the use of XGBoost in business failure prediction but also pioneers the integration of ethical considerations into this field. By demonstrating the transparency and interpretability of the XGBoost model, we address the ethical concerns often associated with the "black box" nature of machine learning algorithms.

Interpretability in machine learning is an area of ongoing research and development. This study contributes to this effort by providing a practical example of how an XGBoost model can be interpreted and explained in the context of business failure prediction.

The paper is structured as follows. Following a brief introduction, we conduct a thorough review of the literature. We then deal with the methodological and empirical part, establishing the definition of business failure and the explanatory variables used in the study. Next, we compile a database composed of companies whose records

meet the specifications required by the objectives of this study, to which we will apply the analytical techniques and machine learning tools. Finally, we present the most relevant conclusions drawn from the study.

2. LITERATURE REVIEW

Studies of business failure that apply prediction models date back to the 1960s, starting with the use of univariate techniques (Beaver, 1966). The ongoing search for methodological improvements to overcome certain flaws and limitations of the pioneering models led to the incorporation of multivariate techniques (Altman, 1968), followed by conditional probability techniques (Ohlson, 1980; Zmijeswski, 1984), and iterative partitioning, before this line of research was eventually reinforced with the use of artificial intelligence, (Frydman et al., 1985; Wilson & Sharda, 1994). This evolution has been accompanied by a gradual but unquestionable improvement in the power, reliability, interpretability and ease of application of the different models developed. There is an extensive bibliography on all of them in the national and international accounting literature (Tascón & Castaño, 2012).

Stef (2021) examines the relationship between institutional quality and firm recovery from financial distress in Central and Eastern Europe. It finds that stronger anticorruption efforts, particularly those implemented before or early in the post-crisis period, are associated with improved financial health of firms in the region.

It is only relatively recently that we have begun to see numerous proposals based on artificial intelligence making inroads into this line of research. The objective, apart from improving on and expanding current detection tools, is to overcome certain restrictions inherent to traditional statistical models, such as linearity, normality and above all the stationarity derived from the use of financial ratios calculated using financial information referring to certain periods. The empirical evidence points to many more variables and determining factors that influence business failure. The most significant proposals of these new artificial intelligence-based methods can be found in a few recent studies, such as that by Momparler et al. (2020), who identified the combinations of factors that lead to bank failure, while Hosaka (2019) applies a convolutional neural network to a sample of Japanese companies to predict bankruptcy. For their part, Le

and Viviani (2018) compare traditional statistical techniques (discriminant analysis and LR) versus ML techniques (artificial neural network, support vector machines, or SVM, and k-nearest neighbors). They conclude by demonstrating the superiority and greater accuracy of artificial neural network and k-nearest neighbor methods when it comes to predicting bank failure. Pozuelo et al. (2018) use the method based on the “Gradient Boosting Machine” (GBM) algorithms to predict the bankruptcy of Spanish firms, confirming the high predictive power of the models based on this method. In addition, Lee and Choi (2013) apply the back-propagation neural network (BNN) method to study bankruptcy in Korean companies, showing that BNN outperforms multivariate analysis in bankruptcy prediction.

Tsai and Cheng (2012) analyze the predictive power of four popular classification techniques. They compare artificial neural networks, decision trees, LR and SVM in the presence of outliers, applying these techniques to Australian, German and Japanese competition datasets. They conclude that the SVM model provides the highest rate of prediction accuracy and performance.

Erdogan (2013) proposes a novel study applying SVM to the analysis of bank failures in a sample of Turkish banks. Her study shows that the SVM can be used as part of an early warning system. Eling and Jia (2018), drawing on a sample of failed European insurance companies, apply state-of-the-art data envelopment analysis, LR and SVM approaches. They conclude that SVM is better than LR.

Zhou et al. (2017) propose an improved method for selecting effective features for predicting the listing statuses of Chinese-listed companies. They use models based on decision trees C4.5 and C5.0 and compare them with other widely used models, demonstrating the efficacy of the proposed feature selection method and decision tree C5.0 model.

Du Jardin's series of papers (2010, 2015, 2016, 2017) laid a foundation for understanding the nuances of financial distress prediction. Perboli and Arabnezhad (2021) highlight the need to understand the contribution of individual models to the overall prediction.

Zięba et al. (2016) develop a predictive model combining various econometric measures using XGBoost, and demonstrate its superior predictive performance over

traditional methods. Jones (2017) concludes that the statistical learning method of XGBoost overcomes the limitation of multiple discriminant analysis and logit models. Momparler et al. (2016) apply the GBM technique to the study of bankruptcy in banking institutions in the eurozone, yielding very satisfactory results in terms of the high predictive power of the model. Carmona et al. (2019) compare the ability of XGBoost, Random Forest and LR to predict bank failure. They report that XGBoost outperforms other models in terms of accuracy and generalization. Jabeur et al. (2021) propose a novel approach to classify categorical data using gradient boosting decision trees, CatBoost. They compare it with eight reference ML models, observing an effective improvement in the power of classification performance compared with other advanced approaches.

Sigrist & Leuenberger (2023) apply machine learning to predict corporate defaults, finding that these methods are more accurate than linear models, especially over longer timeframes. Ben Jabeur et al. (2023) proposes a new bankruptcy prediction model, FS-XGBoost, which uses feature selection to improve the accuracy of the XGBoost algorithm. The results show that FS-XGBoost outperforms seven other machine learning algorithms and three traditional feature selection methods.

We can conclude that practically all artificial intelligence-based models identify financial difficulty better than the "traditional" models. In this study, in addition to helping expand the literature in this field, we seek to reveal the contents of the black box of this methodology and analyze its degree of objectivity in the information provided. It is our view that the autonomy of these systems cannot serve as a pretext for an abdication of responsibility. Therefore, it is necessary to ensure the objectivity of any decision made by incorporating appropriate mechanisms.

The growing prevalence of artificial AI and machine learning ML in financial decision-making has produced a growing interest in the interpretability and ethical dimensions of these models. Researchers and practitioners alike recognize the importance of understanding how these complex algorithms arrive at their conclusions, not only to ensure transparency and accountability but also to address potential biases and ethical concerns.

Several studies have investigated the interpretability of ML models in the context of financial distress prediction. For instance, Carvalho et al. (2019) conducted a comprehensive survey on methods and metrics for machine learning interpretability, highlighting the importance of understanding model behavior and the factors influencing predictions. Similarly, Chen and Guestrin (2016) introduced XGBoost, a scalable tree boosting system, and emphasized the need for interpretability tools to gain insights into the model's decision-making process.

In the field of ethical considerations and ML interpretability, Doshi-Velez and Kim (2017) proposed a roadmap for a rigorous science of interpretability, emphasizing the importance of transparency and accountability in AI systems. They argued that interpretability is not only crucial for understanding model behavior but also for ensuring that AI systems are used ethically and responsibly. Bussmann et al. (2021) explored the potential of Shapley values, a game-theoretic approach, to explain the predictions of complex ML models. Smith and Alvarez (2021) provided a comprehensive overview of XAI techniques, discussing their potential to enhance the transparency and interpretability of ML models in finance. Carmona et al. (2022) apply the XGBoost algorithm to create an interpretable model for predicting business failure.

These studies collectively underscore the growing recognition of the importance of interpretability and ethical considerations in the development and deployment of AI models in finance. As AI continues to play an increasingly significant role in financial decision-making, the need for transparency, accountability, and ethical oversight becomes ever more critical.

In summary, while the literature on business failure prediction is extensive, the specific application of interpretable machine learning models in this field remains relatively unexplored. This study aims to contribute to this growing area of research by demonstrating the transparency and interpretability of the XGBoost model in the context of financial distress prediction.

This study aligns with the burgeoning field of explainable AI (XAI), which seeks to develop machine learning models that are not only accurate but also transparent and understandable. By shedding light on the 'black box' of XGBoost, we contribute to the

advancement of XAI in finance and pave the way for more responsible and trustworthy AI applications.

3. OBJECTIVE, SAMPLE AND EXPLANATORY VARIABLES

3.1. Research objective

The aim of this study is to develop a classification model of companies prone to failure and not prone to failure, by applying ML techniques—in particular, the XGBoost algorithm—to companies' financial data sourced from the ORBIS database.

To do so, we have selected and obtained financial information on active Spanish companies from the 2019 financial year, before the pandemic impacted companies' financial situation. Said financial information includes data from the companies' annual accounts, financial ratios and indicators, and financial strength indicators.

3.2 Indicators of strength. VADIS propensity to go bankrupt indicator.

The ORBIS database offers a series of financial strength indicators that are very useful for analyzing different aspects of companies' viability.

In this study, we have selected the VADIS strength indicator that measures companies' evolution based on a large number of factors (about 15,000 per company). It offers information on the following aspects:

a) Propensity of a company to go bankrupt (P2BB): Measures the probability that a company will file for bankruptcy in the next 18 months.

b) Propensity of a company to be sold (P2BSold). It measures the probability that a company will be sold in the next 18 months.

c) Estimated value of the company's operations (VPI EDV). It estimates the value of the future operations of companies associated with a P2Bsold indicator and is expressed in terms of a confidence interval (that is, it has an upper and lower bound).

The VADIS indicator, associated with the propensity to fail, classifies companies in different grades ranging from 1 (lowest propensity to go bankrupt) to 9 (highest propensity to go bankrupt). The scale used by this indicator is shown in Table 1.

Table 1. Scale of the VADIS financial strength indicator on propensity to go bankrupt

	SCALE OF THE FINANCIAL STRENGTH INDICATOR VADIS P2BB ON PROPENSITY TO GO BANKRUPT:
Value	**The company's risk of bankruptcy in the next 18 months is**
9	more than 10 times the national average
8	between 5 and 10 times the national average
7	between 3 and 5 times the national average
6	between 2 and 3 times the national average
5	between 1 and 2 times the national average
4	between 1/2 the national average and the national average
3	between 1/5 and 1/2 of the national average
2	between 1/10 and 1/5 of the national average
1	below 1/10 of the national average

Source: Own elaboration based on information from the ORBIS database (2021).

The scores of this indicator are updated monthly if the companies in question have available, recent and detailed financial data in the ORBIS database, the quality and coverage of which is essential for the model.

However, the classification offered by the indicator is not transparent to the user, since the underlying calculations and reasons are not known in any detail. Therefore, in this paper we seek to elucidate the factors that characterize the companies classified by the VADIS indicator as prone to business failure compared to those that are not, using the algorithm XGBoost to do so.

3.3 Sample selection

From the ORBIS database, we have sourced financial information on active Spanish companies for the financial year 2019, from all sectors of activity except those in the financial field.

We have thus obtained data and financial ratios for 42,247 companies, in addition to information on the VADIS financial strength indicator. Based on that, we have identified

two groups of companies, one with a non-high propensity to fail and the other with a very high propensity to fail.

To that end, we consider companies that have a high value of the VADIS financial strength indicator (grades 9 and 8) as being highly prone to failure, and those that have a low value in said indicator (1 to 5) as being not highly prone to failure; to prevent any interference we leave out those assigned an intermediate value in said indicator (grades 6 and 7).

Using this criterion, we have identified a total of 1,809 companies with a high propensity to fail and we have randomly selected 1,800 companies whose propensity to fail is not high, as detailed in Table 2. The sample of companies is balanced in terms of the proportions of the two groups, which is important for the correct application of the algorithms.

Table 2. Sample selection

VADIS Classification	1	2	3	4	5	6	7	8	9
Companies	2,159	6,047	13,926	5,528	6,348	2,611	1,867	1,334	475
Selection	34,008→randomly→1,800					0	0	1,809	
Propensity to fail	Not high							High	

Source: Own elaboration based on data sourced from the ORBIS database (2021).

As a last step, to test the predictive or classificatory power of the model and its degree of generalization, 20% of the observations (714 companies) have been reserved as an independent test sample, to check the results obtained in the training or estimation sample, made up of the remaining 80% of the observations (2,895 companies). All the performance and accuracy indicators of the model obtained through the XGBoost ML algorithm included in this study have been obtained from the independent test sample.

3.3. Selection and definition of explanatory variables

Among the most relevant aspects when building models for predicting or classifying business failure is the selection of independent variables, which in our case are fundamentally economic/financial ratios.

Since there is no specific theory establishing the selection process for these variables—which constitutes a limitation when modeling business failure—we have tried to incorporate the experience of authors such as Tascón and Castaño (2012), Tian and Yu (2017), Mselmi et al. (2017) and Khoja et al. (2019), who analyze in detail the variables used in the main studies on business failure prediction.

The list and descriptions of the variables initially used, separated by category, are shown in Table 3. We have also considered in the election of the variables the study of Liang et al. (2016) which investigates the effectiveness of combining different types of ratios in bankruptcy prediction.

Table 3. Explanatory variables

KEY	VARIABLES
PROFITABILITY RATIOS	
RSHF	ROE (base EBT) (%) = (EBT / equity) * 100
RCEM	ROCE (base EBT) (%) = [(EBT + interest paid) / (Permanent sources of financing)] * 100
RTAS	ROA (base EBT) (%) = (EBT/ Total Assets)*100
ROE	ROE (base Net income) (%) = (Net income /Equity) * 100
ROCE	ROCE (base Net income) (%) =[(EBT) / Permanent sources of finance] * 100
ROA	ROA (base Net income) (%) = (Net income / Total Assets) * 100
PRMA	Profit margin (%) = (EBT/Operating revenue) * 100
ETMA	EBITDA margin % = (EBITDA /Operating revenue) * 100
EBMA	EBIT margin (%) = (EBIT/Operating revenue)
CFOP	Operating cash flow (%) = (Cash flow / Operating revenue) * 100
OPERATIONAL RATIOS	
NAT	Net asset turnover = Operating revenue/ (Perm. sources)
COLL	Collection Period (days) = (Debtors / Operating revenue) * 360
CRPE	Credit Period (days) = (Creditors / Operating revenue) * 360
STRUCTURAL RATIOS	
CURR	Current Ratio = (Current assets/ Current Liabilities)
LIQR	Liquidity Ratio = (Current Assets - Stocks) / Current liabilities
SOLL	Solvency Ratio (%) = (Equity /Total Assets) * 100
GEAR	Indebtedness Ratio (%) = [(Non-current liabilities + Loans) / equity] * 100

RCD	Debt Quality Ratio = Current Liabilities/Total Liabilities
PER EMPLOYEE RATIOS	
PPE	Profit per employee = EBT/ no. employees
TPR	Operating revenue per employee = Operating revenue / no. employees
SCT	(Staff costs/Operating revenue) * 100
ACE	Average cost of employees = Staff costs / no. employees
SFPE	Shareholder funds per employee = equity / no. employees
WCPE	Working capital per employee = Working capital / no. employees
TAPE	Total assets per employee = Total assets / no. employees

Note: Key based on the reference that appears in the ORBIS database (2021).

Source: Own elaboration based on information sourced from the ORBIS database (2021).

The 25 explanatory variables initially selected, having previously discarded some for presenting too many missing values, have been used in the application of the XGBoost algorithm to try to identify the ones that are most relevant when it comes to determining which companies are most prone to business failure, as we describe in the following section.

4. METHODOLOGY

The aim of this study is to develop a classification model using ML techniques; specifically, we make use of the XGBoost algorithm. As noted above, the ORBIS database provides us with the VADIS financial strength indicator, from which we have identified a total of 1,800 companies whose propensity to fail can be categorized as "not high" and 1,809 companies with a "very high" propensity to fail, together totaling 3,609 companies. It is very important that the model developed is not overfitted and is generalizable to new data. Chen and Guestrin (2016) highlight that XGBoost is an efficient and scalable implementation of the gradient boosting algorithm proposed by Friedman (Friedman, 2001; Friedman et al., 2000; Friedman, 2002). One of the elements that characterizes algorithms based on gradient boosting is the fact that decision trees are created sequentially, so that each new tree is built to correct the errors of the previous ones.

The use of this methodology produces excellent prediction results, as has been confirmed in the recent studies by Climent et al. (2019) and Carmona et al. (2019), aimed at predicting banking failure in Europe and the US respectively.

In fact, the XGBoost algorithm produces a set of simple models based on decision trees, such that summing them together considerably improves the predictive power of the individual decision trees. Another feature of an XGBoost model is that it incorporates into the construction of the decision trees something known in ML as regularization. It entails the implementation of a mechanism to control the weight of the variables in the final model, which is an element that sets it apart from similar algorithms also based on gradient boosting, and significantly reduces the possibility of developing models that cannot be generalized to independent data.

As noted earlier, we have a total of 3,609 observations of companies, which constitutes the initial sample. To correctly apply the ML techniques, we have divided these observations into two groups: the first forms the training sample (2,895 observations, approximately 80%), and the second the test sample (714 observations, approximately 20%). We will use the first group or training sample to train the XGBoost algorithm. Then, having identified the best of the models, we will use the test sample, which is an independent sample and has not been used to train the algorithm, to make sure that the results are valid and therefore generalizable.

Obtaining the best XGBoost model for the available data set requires the identification of the optimal hyperparameter values. Doing so entails an iterative process with the different combinations of values that these hyperparameters can take to single out the best of all these combinations, thus allowing the construction of a model with high predictive power. Below, we outline to the main hyperparameters of the XGBoost algorithm:

- The number of rounds or passes over the data that the algorithm needs. In each round, an individual decision tree is constructed. There is generally a limit marking the number of passes after which the resulting model does not improve.

- Maximum depth of the individual decision trees. The lower this value, the lower the ability of the trees to capture the features of the data, but a high value can lead to overfitting problems.

- Learning rate or contribution of each tree in the growth of the model. It determines how quickly the algorithm incorporates or adapts the contribution of each tree to the model under construction in each iteration. Small values are recommended to ensure slow growth and thus avoid overfitting problems.

- Gamma or minimum loss reduction for the algorithm to make a further partition. When the reduction of the loss function is below the indicated value, no more partitioning of the decision trees will be made. This parameter controls the regularization of the model: the higher the specified value, the higher the regularization.

- Ratio of the percentage of the variables and observations involved in the construction of each tree.

- Minimum number of observations needed in each terminal node. The higher this number, the more conservative the algorithm will be, in the sense that it will be more flexible and robust to overfitting problems.

To identify the best possible combination of all these hyperparameters, ML practitioners commonly make use of cross-validation techniques. Specifically, the decision was made to perform cross-validation with 10 subsamples (k = 10), so the training process for the XGBoost algorithm consists of fitting 10 models on 9 different subsamples (each with 90% of the observations), so that the final goodness of fit is estimated from the mean of the results obtained in the remaining 10% of each of the 10 subsamples. Furthermore, to identify a good combination of hyperparameters that yields a very robust model with high predictive power, we rely on what is known as random search; that is, from all the feasible configurations in the search space specified, 25 models have been developed, corresponding to 25 possible random combinations, or in actual fact 250 (25 x 10) due to the cross-validation with 10 subsamples.

As indicated above, all the available data on the companies considered has been divided into two groups. On the one hand, we have the training sample with 80% of the observations, which has been used to train the XGBoost algorithm and identify a good

combination of hyperparameters that yields a final model with high predictive power. On the other hand, we have the remaining 20%, which as an independent test sample has been used to validate the final model and determine its performance metrics.

To implement the XGBoost algorithm, we use the statistical programming environment R (R Core Team, 2022) and the H2O library (LeDell et al., 2022).

4.1. XGBoost model with all the variables

Following the methodology described above, we have constructed an XGBoost model that incorporates all the variables listed in Table 3 to identify the most relevant variables and the ones that are therefore crucial for classifying a company with a high probability of facing a situation of bankruptcy in the near future. As is common in ML classification scenarios, we have considered the AUC (area under the curve) indicator to identify the best of the models fitted according to the different combinations of hyperparameters. This metric compares the false positive rate with the true positive rate. This area can intuitively be seen to represent the probability that a randomly selected company with a high propensity to go bankrupt has a higher probability of actually being classified by the final model as highly prone to bankruptcy. The maximum value that this indicator can take is 1 and the higher the value, the better the resulting model.

As indicated above, we have constructed 25 primary XGBoost models, corresponding to 25 possible combinations of hyperparameters, and their corresponding cross-validation models. For the final model we have selected the one with the highest AUC value on the cross-validation samples, which registered a value of 0.99, very close to the maximum value of 1; this model achieved a global accuracy of 96%, also on the cross-validation samples. In Table 4 we present the values taken by the hyperparameters for this model.

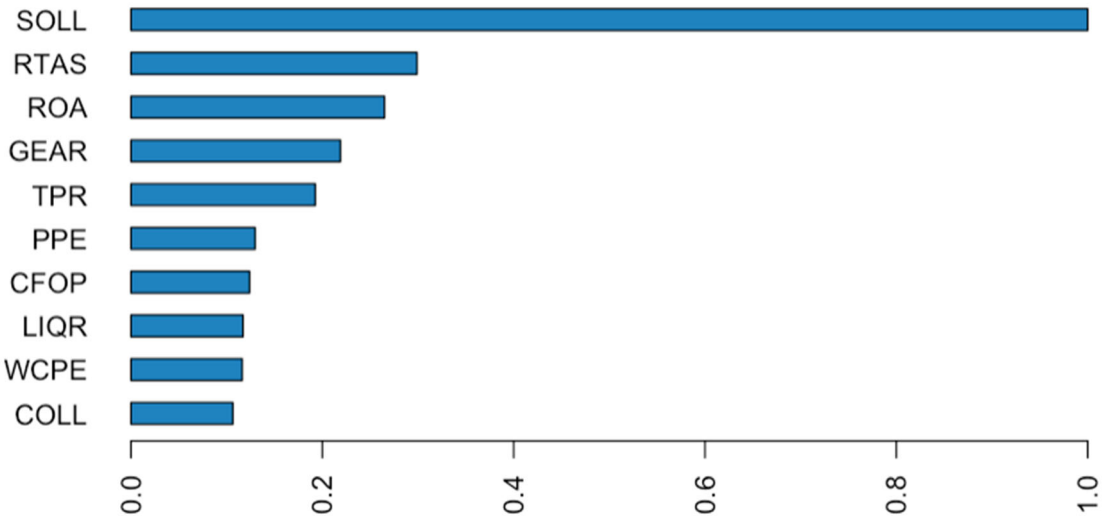
Table 4. Optimal values for the hyperparameters of the *XGBoost* model.

Number of passes of the algorithm	60
Maximum depth	12
Learning rate	0.19

Gamma	0.001
Percentage of variables	46%
Percentage of observations	70%
Minimum observations terminal node	2

Figure 1 shows the variables that are most important in the construction of this XGBoost model for classifying companies that have a very high propensity to go bankrupt in the following 18 months. For this type of prediction model based on the gradient boosting algorithm, and in order to identify the most important variables, we consider the gain that each variable incorporates into each of the trees constructed and then calculate the average across all the variables to obtain a global view of the model as a whole (Chen & Guestrin, 2016). The first four variables in Figure 1 are those of the greatest importance: SOLL, RTAS, ROA and GEAR.

Figure 1. Most relevant variables in the *XGBoost* classification model



Next, we use partial dependence plots to represent the relationship between the most relevant variables and a company's propensity to fail. These are global interpretation measures that help us to understand the relationships between the inputs of the model and the output or dependent variable (Naketin & Knoll, 2013; Hall & Gill, 2019). These types of depictions offer an understanding of the process that a model follows to make predictions according to the importance of the detected variables and the influence

they exert on the structure of the model. These dependence plots show the response of the model as a function of one of the independent variables, taken as the mean of all the observations.

Figure 2 shows the partial dependence plots for the four variables identified as being the most relevant. Additionally, Figure 3 shows the joint representation of the distribution and box plot for the most important variable, SOLL, both for high propensity and non-high propensity companies.

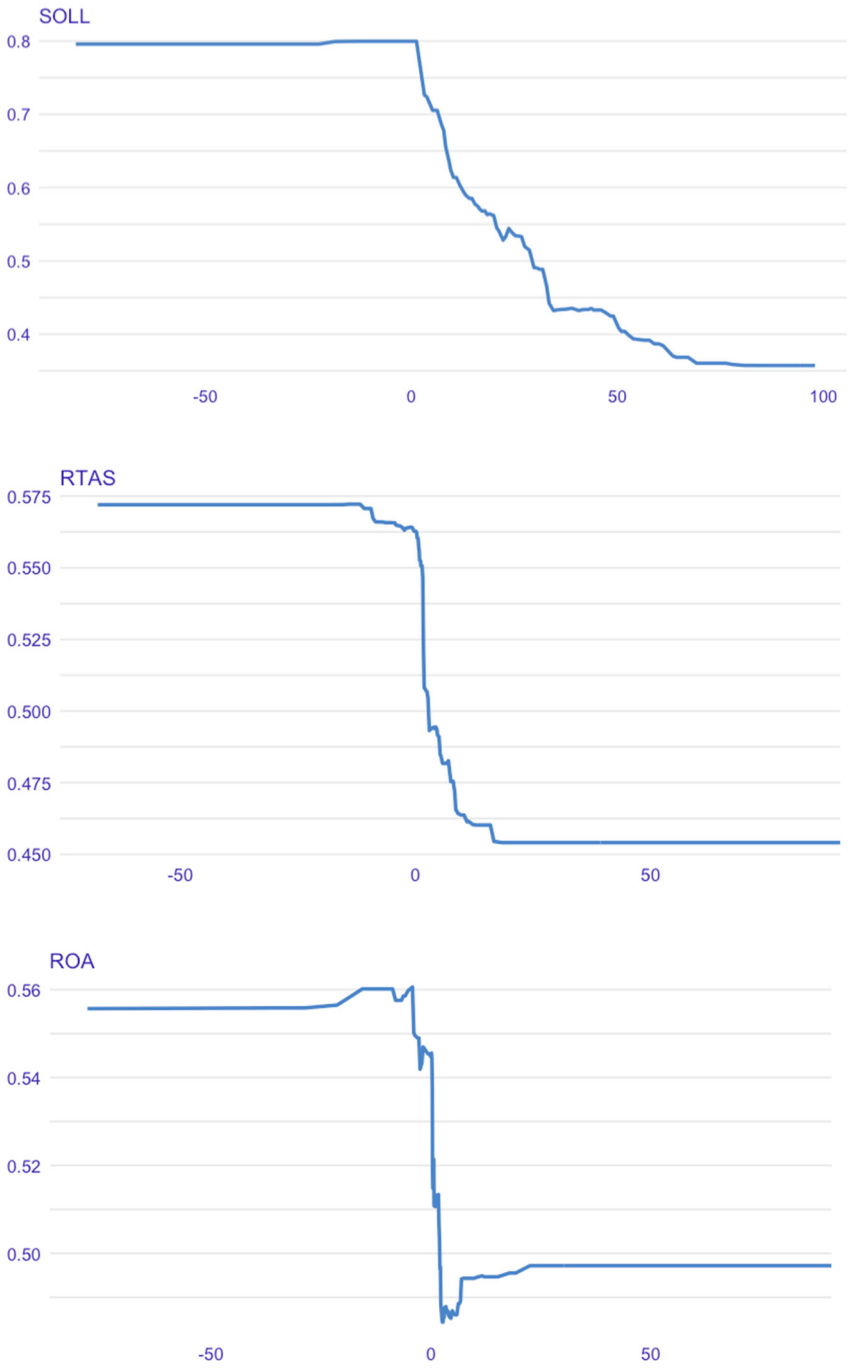
The partial dependence plots of the four most important variables, depicted in Figure 2, illustrate the marginal effect of each variable separately on the probability that a company can be considered prone to bankruptcy. The x-axis shows the values that the variables can take, and the y-axis shows the probability of the company going bankrupt.

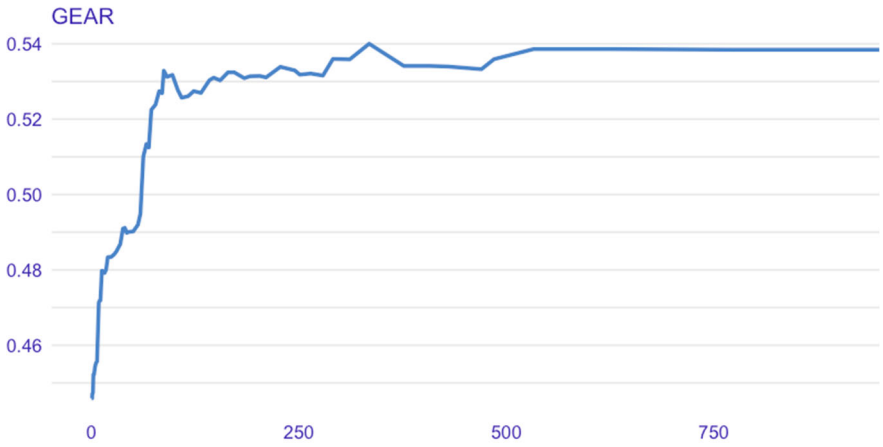
The downward curve of the graph corresponding to the SOLL variable indicates that as the solvency of the company increases, its propensity to fail decreases. This result is entirely consistent since if the company can cope with its debts, there will logically be a smaller possibility of it getting into a situation of financial difficulty that leads to failure.

The downward slopes of the RTAS and ROA variables in Figure 2 indicate that the higher their value, the lower the propensity to fail. These variables increase when EBT or Net income, respectively, rises relative to the company's assets, with it being reasonable to suppose that when a company is more profitable, its propensity to fail is reduced.

Finally, taking into account the upward slope of the GEAR variable in Figure 2, which analyzes the indebtedness of the company, it can be seen that as its value increases, the probability of failure also rises. The gearing ratio increases with a rise in external financing, so a high amount would reflect excessive indebtedness relative to internal financing, which could give rise to problems in the company. Accordingly, it is logical that high values of this variable are associated with the company having a higher propensity to fail.

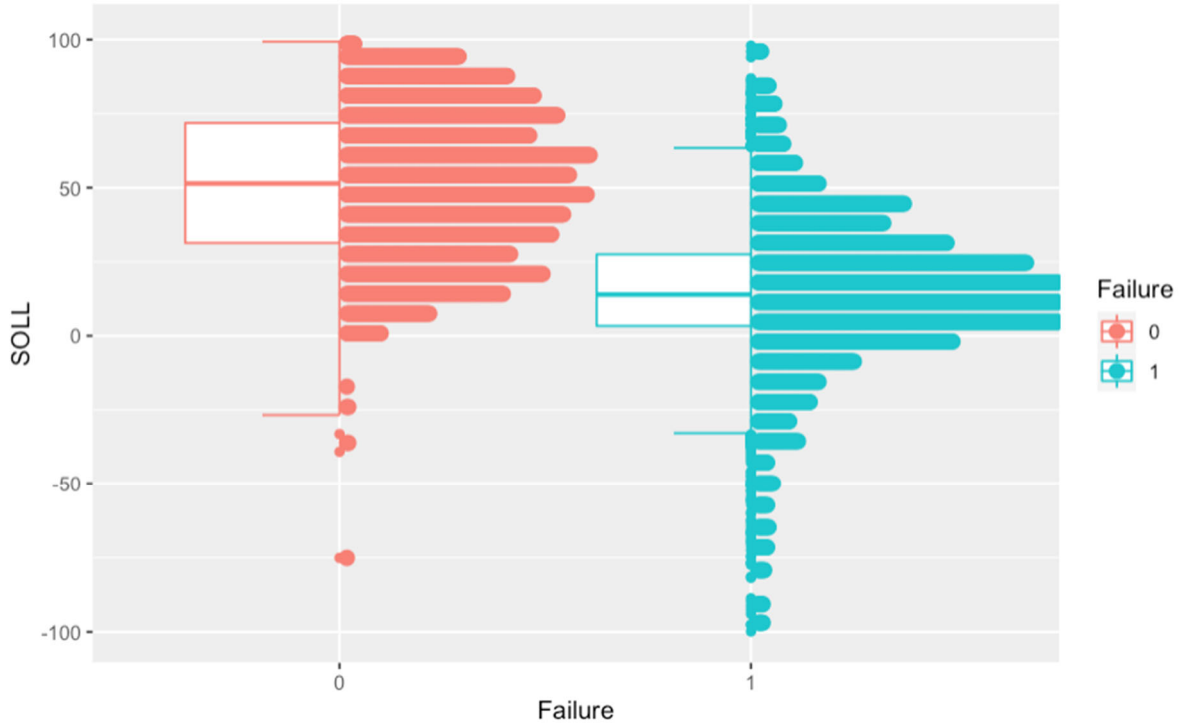
Figure 2. Partial dependence plots





NOTE: the *y*-axis represents the probability of the propensity to bankruptcy

Figure 3. Joint representation of the distribution and box diagram of the SOLL variable



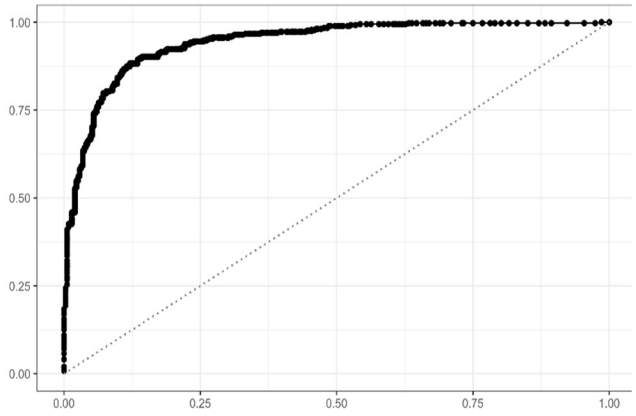
4.2. Validation of the model on the test sample

Below, we check the predictive power of this model on an independent data sample or test sample. To do so, for this test sample (see Figure 4) we have taken the prediction

values of the XGBoost classification model fitted and compared them with the actual values, registering a global accuracy level of 88% and an AUC value of 0.943—very close to the maximum value of 1. Therefore, in view of these results, we can conclude that the XGBoost model identified does not present overfitting problems and can be generalized and applied to independent samples.

Figure 4. Results of the *XGBoost* model on the test sample

AUC (area under the curve)



NOTES: AUC = 0.943. Accuracy = 88.24%.

x axis: false positive rate.

y axis: true positive rate.

Confusion matrix

Actual values		Predicted values: Propensity to go bankruptcy: No	Predicted values: Propensity to go bankruptcy: Yes	Accuracy
Bankruptcy	No	**296**	52	85.06%
	Yes	32	**334**	91.26%
Total				88.24%

NOTES:

Results for the best cut-off point of 0.24

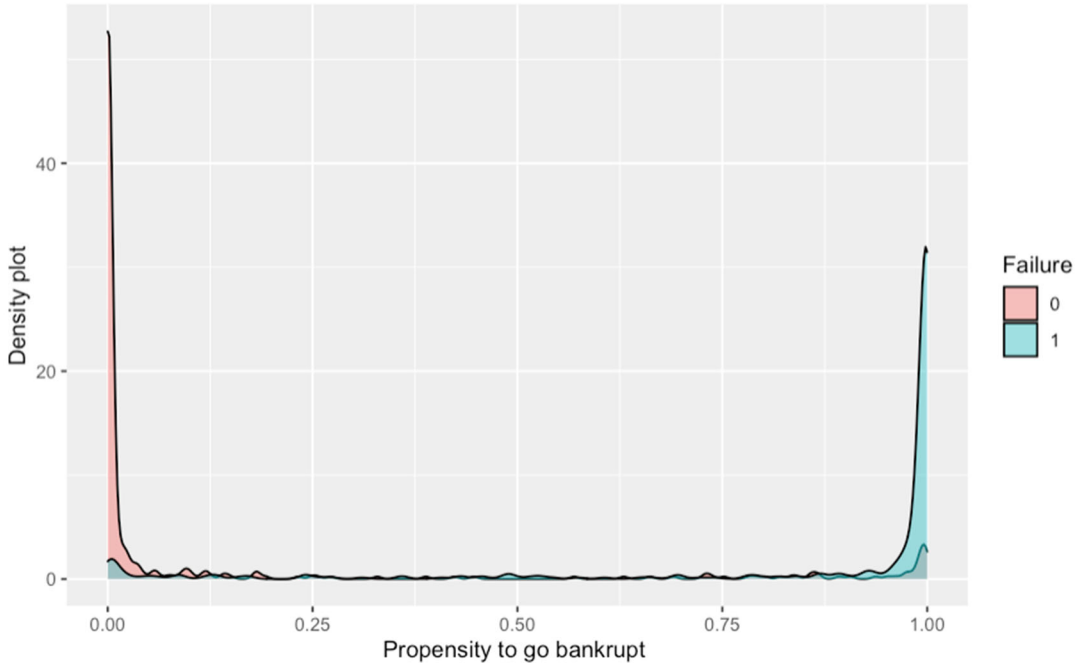
N = 714

Furthermore, Figure 5 shows the density function of the probabilities of going bankrupt in the following months generated by the model, separating companies with an actual high propensity of bankruptcy and those with an actual low propensity according to the VADIS rating, and applied to the data of the test sample. The XGBoost model estimated can very effectively separate companies with a high propensity to bankruptcy and those that do not have a high propensity, such that we can draw an imaginary vertical line that clearly differentiates these two groups.

In short, the high performance values obtained on the test set, in terms of the model's overall accuracy and AUC, along with the graphical representations of failure

probabilities, confirm that the results achieved with the XGBoost algorithm on the training sample can be extrapolated or generalized to independent samples. The use of cross-validation, an ML technique, has largely contributed to the identification of a robust model.

Figure 5. Density plots of the distribution of the probability of failure of the *XGBoost* model (test sample)



NOTES: 1.00 indicates a high propensity to go bankrupt and 0.00 a non-high propensity to go bankrupt

4.3. Comparing XGBoost with two other classification models

Having fitted a model using the XGBoost algorithm, which has demonstrated strong predictive capabilities in classifying companies based on their financial health (grouping companies as either prone to failure or not prone to failure), we now turn our attention to a comparative analysis. This analysis will involve two distinct approaches: a classical logistic regression and a modern deep learning model. The former represents a well-established statistical technique frequently employed in classification tasks, while the latter is a cutting-edge approach within the machine learning domain

known for its capacity to learn complex patterns from data. By comparing the performance of these models, we aim to gain insights into the relative strengths and weaknesses of each approach in the context of corporate failure prediction.

Logistic regression is a widely used traditional statistical method for binary classification problems, where the goal is to predict the probability of an event occurring (in this case, prone to failure) based on a set of predictor variables (such as financial ratios). It models the relationship between the predictors and the outcome using a logistic function, which produces a probability estimate ranging from 0 to 1.

Mathematically, the logistic function is represented as:

$$P(Y=1|X) = 1 / (1 + \exp(-(\beta0 + \beta1X1 + \beta2X2 + ... + \betapXp)))$$

where:

- $P(Y=1|X)$ is the probability of the event ($Y=1$) occurring given the predictor values (X).
- $\beta0$ is the intercept term.
- $\beta1, \beta2, ..., \betap$ are the coefficients associated with the predictors $X1, X2, ..., Xp$, respectively.
- exp() is the exponential function.

The model identifies the optimal values of the coefficients (β) that maximize the likelihood of observing the data. This is typically done using maximum likelihood estimation (MLE) or other optimization techniques.

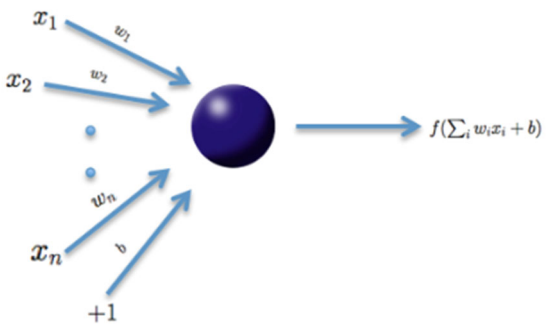
The linear combination of predictors ($\beta0 + \beta1X1 + \beta2X2 + ... + \betapXp$) is often referred to as the log-odds or logit. The logistic function transforms this logit into a probability value.

Once the model is trained, it can be used to predict the probability of the outcome for new observations based on their predictor values.

On the other hand, deep learning, a subfield of machine learning, utilizes artificial neural networks with multiple layers to learn complex patterns and representations from data (Heaton et al., 2018). In contrast to traditional models, deep learning models automatically extract features and hierarchies of features, reducing the need for manual feature engineering (LeCun et al., 2015).

The fundamental building block of deep learning is the artificial neuron, which receives inputs, applies weights and biases, and produces an output through an activation function. These neurons are organized into interconnected layers, forming a network. The input layer receives raw data, hidden layers process and transform the information, and the output layer produces the final prediction.

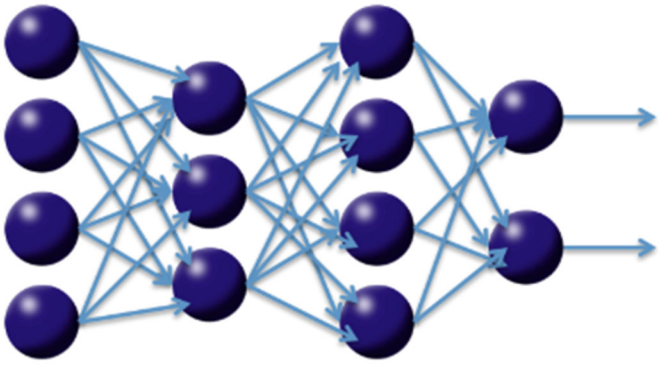
Figure 6. Neuron from a hidden layer



Source: Candell & Ledell (2018)

Deep learning models, such as multi-layer feedforward neural networks (Figure 7), are trained using large datasets and backpropagation, an algorithm that adjusts the weights and biases of the network to minimize the difference between predicted and actual outcomes. Each neuron has a weight assigned (w1, w2, w3... wn) to each of its inputs, which are modified until they reach optimal values during the neural network's learning process. The impulses or information from each neuron that pass to the next layer are obtained through activation functions (Candel & Ledell, 2018). This iterative process allows the model to learn complex relationships and representations within the data. For a more comprehensive discussion of the inner workings of deep learning models, including their architecture, training algorithms, and applications, readers are encouraged to consult Romero et al. (2021).

Figure 7. Multi-layer feedforward neural networks



Source: Candell & Ledell (2018)

In the context of corporate failure prediction, deep learning models can handle high-dimensional data and capture non-linear relationships, which makes them a promising alternative to traditional statistical models like logistic regression (Barboza, Kimura, & Altman, 2017). For our analysis, we will employ a deep learning model specifically designed for classification tasks. This model will consist of multiple hidden layers with non-linear activation functions, enabling it to learn complex representations of the input data.

Table 5. Performance metrics of the three fitted classification models on test set

Model	Accuracy	True positive rate	True negative rate	AUC	f1-score
XGBoost	88.24%	91.26%	85.06%	0.94	0.87
Logistic	85.15%	85.80%	84.48%	0.92	0.85
Deep Learning	86.27%	93.99%	78.14%	0.93	0.86

By comparing the performance of these two models to that of the fitted XGBoost, we aim to assess the potential advantages of the latter in predicting corporate failure. Table 5 summarizes the key performance metrics for these two models on the test set, along with the results of the initially considered XGBoost algorithm. A comparison of these results reveals that the XGBoost model consistently performs better both the

logistic regression and deep learning models across multiple metrics (e.g., accuracy, F1-score, or AUC).

Figure 8. Reverse cumulative distribution of residuals on test set

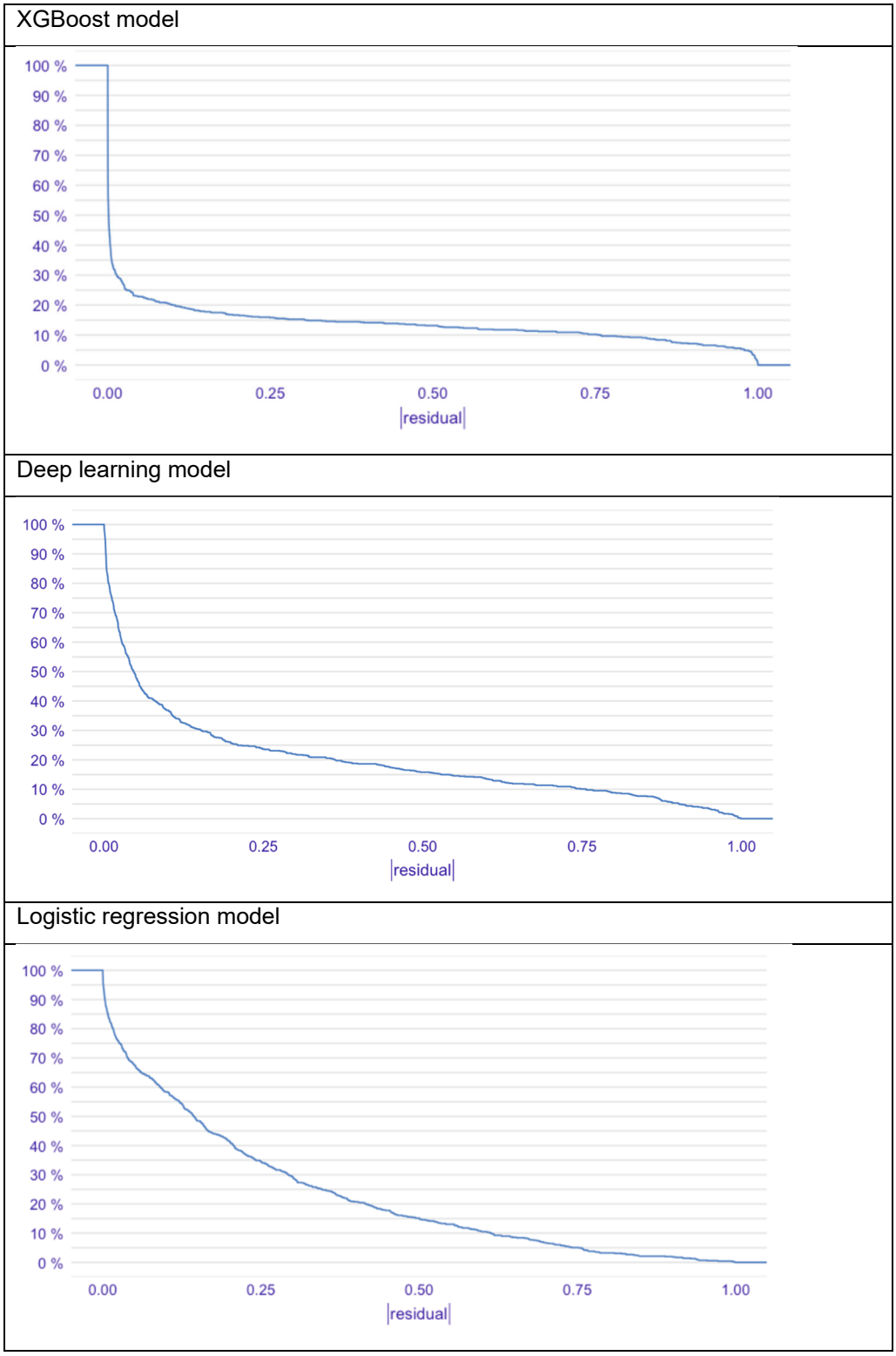


Figure 8 presents the reverse cumulative distribution of residuals absolute values for XGBoost, logistic regression and deep learning models on the test data, where the lower the area under the curve the better. The results indicate that XGBoost exhibits a superior fit, as its residuals generally have lower values. This suggests that XGBoost is more effective at capturing the underlying patterns in the data and making accurate predictions.

In summary, the comparative analysis of XGBoost, logistic regression, and deep learning models for corporate failure prediction reveals the superior performance of XGBoost. This model consistently performs better than the others in terms of accuracy, F1-score, and AUC. Additionally, the analysis of residual distributions demonstrates that XGBoost achieves a better fit to the data, suggesting a greater ability to capture the underlying factors contributing to financial distress. These findings underscore the potential of ensemble learning methods, such as XGBoost, for addressing the complexities of corporate failure prediction

4.4. Robustness analysis

Traditional statistical analysis often relies on strong assumptions about the data, such as normality and homoscedasticity. It focuses on inference and hypothesis testing, aiming to draw conclusions about populations based on samples. In contrast, machine learning prioritizes prediction and generalization to new data. It often deals with complex, high-dimensional data where traditional assumptions may not hold (Cho et al, 2024).

Robustness testing in ML is crucial because it helps assess a model's ability to perform well under various conditions. It ensures that a model is not overly sensitive to small changes in the input data and can generalize to real-world scenarios (Ben Braiek & Khomh, 2024).

XGBoost is inherently a robust algorithm due to its built-in regularization, ensemble learning, and cross-validation mechanisms. Hence, XGBoost already incorporates several mechanisms that promote robustness, such as:

- Regularization: Prevents overfitting and improves generalization.

- Ensemble learning: Combines the predictions of multiple trees to reduce variance and errors.
- Cross-validation: Assesses model performance on different subsets of the data to ensure generalization.

These features make XGBoost inherently more robust than some other ML algorithms and traditional statistic techniques.

In summary, while robustness testing is generally important in ML, its necessity and specific methods can vary depending on the algorithm and the application. In the case of XGBoost, it is inherently more robust than some other ML algorithms due to its ensemble nature and regularization techniques. It combines the predictions of multiple trees, which helps to reduce overfitting and improve generalization. Additionally, the regularization penalizes complex models and enhances robustness (Kharkar, 2022).

We further investigated the robustness of our findings by comparing the most important variables identified by the XGBoost model (Figure 1) with those derived from the two additional models: logistic regression (Figure 9) and deep learning (Figure 10). The results demonstrate that, in general, the three models consistently identified the same variables as the most relevant features, supporting the robustness of our initial feature selection.

Figure 9. Most relevant variables in the *logistic* model

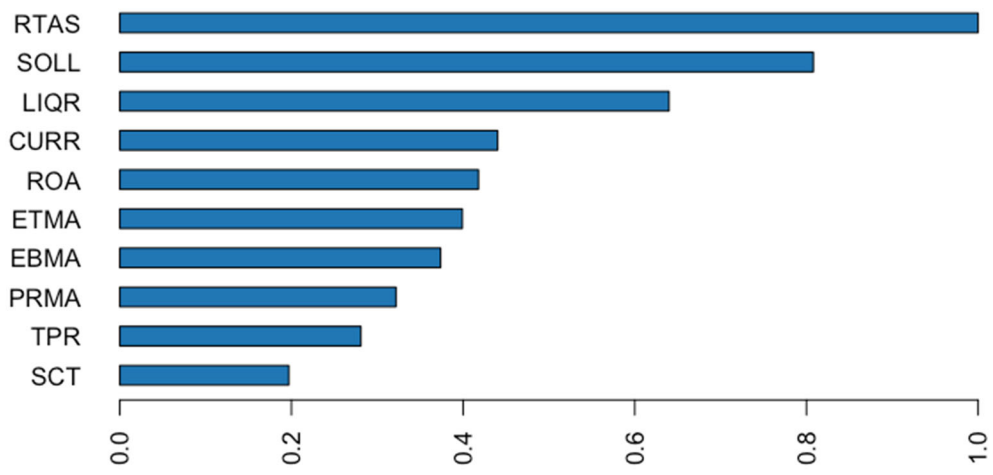
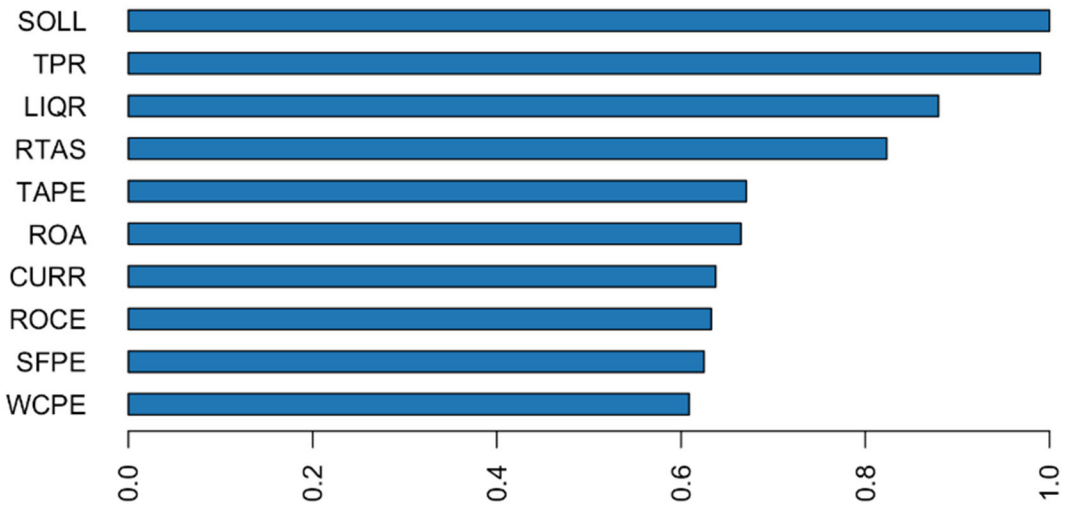


Figure 10. Most relevant variables in the *Deep Learning* model



5. TRANSPARENCY OF XGBOOST FINANCIAL DISTRESS PREDICTION AT THE INDIVIDUAL LEVEL: A LOCAL INTERPRETATION

In the final part of our study, we move beyond the usual focus on machine learning and explore how to make the XGBoost model more transparent, where interpretability becomes crucial. We explain how we can fully understand the model by showing how it determines the effect of each variable on every observation within our dataset (at the level of individual instances). To illustrate this concept, we select two instances where the model assigns a high probability of bankruptcy (as depicted in Figure 11). Conversely, we also consider two instances with a low probability of bankruptcy (as depicted in the final two graphs of Figure 12).

These visual representations constitute a critical aspect of what is commonly known as the "local interpretation" of a model, a methodology that provides profound insights into the workings of a machine learning model and its predictions for individual observations (Hall and Gill, 2019). In our context, such an approach proves invaluable, as it allows us to analyze the reasons behind the XGBoost algorithm's classification of a particular company as being susceptible to a business failure This process is embedded in the examination of the independent variables and their respective values for the company in question. Through this exercise, we bring to the forefront the ethical implications of our findings, rendering the individual predictions entirely transparent and dispelling the obscurity often associated with the "black box."

In essence, our pursue to clarify the inner workings of the XGBoost model is driven by the necessity to understand the complexities of its predictions. We know that researchers and financial analysts are very interested in understanding why the XGBoost model predicts a specific company is at risk of financial trouble. By figuring out which factors have the biggest impact on these predictions, we can better explain how the model works. This is important for ensuring transparency and accountability, which are key ethical considerations when using AI models in finance.

Therefore, as we attempt to reveal the mechanisms underlying the XGBoost model, it becomes evident that there is a pressing need to identify the variables that exert the most significant influence on predictions for individual companies. This attempt not only enhances the comprehensibility of our predictive model but also serves as an indication to our commitment to transparency and ethical responsibility. In essence, by opening the "black box" and shedding light on the inner workings of our model, we aspire to contribute to a deeper understanding of both the quantitative aspects of machine learning and the ethical implications of its applications within the context of financial analysis.

Biecek and Burzykowski highlight several approaches for interpreting complex, black-box models on an instance level. Among these, two methods have gained particular notoriety: break-down plots and Shapley values. However, they differ in their methodology and underlying principles.

- Break-down plots decompose model predictions into contributions from different variables. They work by identifying the variables that, when changed, have the most significant impact on the prediction. This is done in a greedy way, selecting one variable at a time to minimize the difference between the model's prediction and the prediction after the variable is "relaxed" (allowed to vary according to its distribution). Break-down plots are relatively easy to interpret and computationally efficient.
- Shapley values are a concept from game theory that has been adapted to explain machine learning models. They attribute the prediction of a model to different features by considering all possible combinations of features and calculating the average marginal contribution of each feature across all combinations. Shapley values have desirable theoretical properties, such as fairness and consistency, but they can be computationally expensive to calculate, especially for complex models.

In our study, we opted for break-down plots due to their computational efficiency and ease of interpretation. Break-down plots provide a clear, visual representation of how each feature contributes to a model's prediction, making them particularly useful for communicating results to a public who may not have a deep understanding of machine learning theory. Given these advantages of break-down plots, the remainder of this section of the paper will focus on the break-down plots methodology for model explanation

When analyzing a model's prediction for a specific case, it is natural to wonder which factors are most influential. While no single solution exists, break-down plots offer a useful approach. These plots decompose the model's prediction into contributions from different explanatory variables, providing insights into the factors driving the result. This method is similar to the algorithm explanation introduced by Robnik-Šikonja and Kononenko (2008).

We assume that prediction *f(x)* is an approximation of the expected value of the dependent variable Y given values of explanatory variables *x*. The underlying idea of break-down plots is to capture the contribution of an explanatory variable to the model's prediction by computing the shift in the expected value of Y, while fixing the values of other variables[1].

In machine learning, break-down plots provide a very simple way of summarizing the effect of each of the predictors on the dependent variable of the model. Thus, for a particular observation, the probability that a company is prone to bankruptcy is decomposed into the different impacts corresponding to each of the variables of the model. The Figures 11 and 12 contain these break-down plots.

In these figures it can be seen that the intercept gives a general probability of bankruptcy of 50.90% (0.509), which is the value obtained without taking into account the predictive power of the predictors, and represents the proportion of companies classified by the XGBoost model as having a high risk of bankruptcy compared to the total of the companies included in the database—the mean of the predictions. For

[1] For a comprehensive discussion on the theory and applications of break-down plots, including the mathematical foundations and interpretations, readers are encouraged to consult Staniak and Biecek (2019).

example, for the first company in Figure 6, the prediction of failure yields a probability of 100.00% (1.00), which can be decomposed to reflect the influence of the most relevant variables as follows:

+ 0.509: Intercept

+ 0.297: RTAS = -39.13 (Economic profitability) [now the prediction is 0.806]

+ 0.002: ROA = -29.87 (Net income/Total Assets) [now the prediction is 0.808]

+ 0.190: SOLL = 17.76 (Solvency Ratio) [now the prediction is 0.998]

+ 0.002: GEAR = 232.2 (Indebtedness Ratio) [now the prediction is 1.000 = 100%]

Therefore, break-down plots illustrate how each individual factor's contribution modifies the average prediction of the model, resulting in the actual prediction for a specific case or observation.

In our case, these graphs, provided as an example, reveal that low values of solvency and profitability and high values of indebtedness push the prediction model to assign high values to the probability that a company might go bankrupt in the coming months. Conversely, values at the opposite extreme yield very low figures in the estimation of the propensity to go bankrupt.

Figure 11. Break-down plot for two observations with a very high probability of failure according to the *XGBoost* model

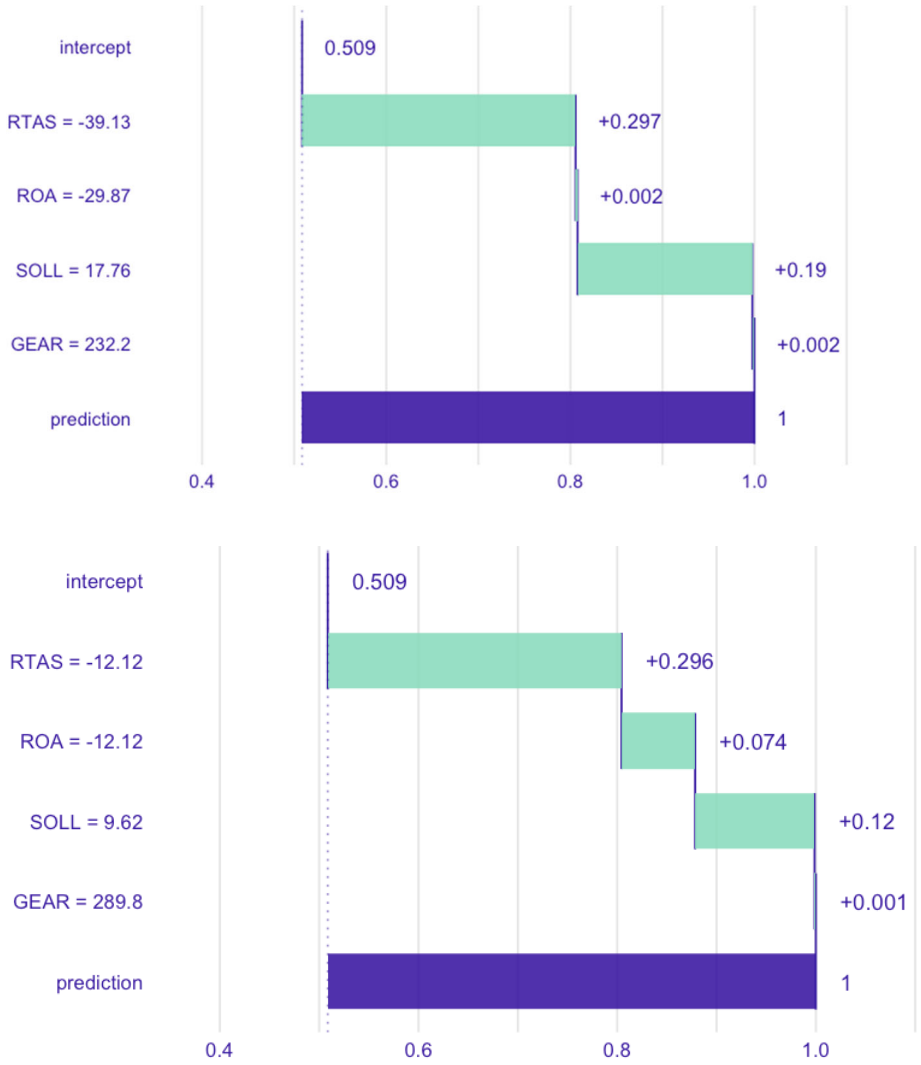


Figure 12. Break-down plot for two observations with a very low probability of failure according to the XGBoost model

In summary, this kind of representation helps to expand on the importance of transparency and interpretability in ensuring fairness, accountability, and avoiding discriminatory practices. To avoid potential biases in the results of the algorithms, we encourage the need for human oversight in the decision-making process.

6. CONCLUSIONS

In this study, we emphasize the power of the Extreme Gradient Boosting (XGBoost) machine learning (ML) algorithm to construct a robust predictive model for identifying companies at risk of bankruptcy. More importantly, our research transcends the inherent opacity of complex ML models by offering a comprehensive and transparent interpretation of their inner workings.

Traditionally, such models have been perceived as "black boxes" due to their obscured internal mechanisms. In our pursuit of understanding, we undertook the ambitious task of explaining these black boxes. Through rigorous analysis, we pinpointed the most influential explanatory variables and clarified their impact on predictions at the individual level. This attempt goes beyond conventional machine learning by shining a light on the often-hidden aspects of model decision-making.

Our study opens these algorithmic black boxes, revealing the critical indicators that characterize a company's propensity for business failure. We utilized the VADIS P2BB financial strength indicator to distinguish between companies susceptible to failure and those resilient to it, based on high or low ratings. Our results underscored the collective significance of solvency, indebtedness, and profitability categories, achieving a robust 88% accuracy rate on an independent test sample.

Specifically, our model identified SOLL, RTAS, ROA, and GEAR as pivotal variables. SOLL gauges a company's solvency, with higher values indicating reduced bankruptcy risk, reinforcing the importance of financial stability. Conversely, the GEAR variable, reflecting indebtedness, suggested that increased indebtedness correlated with heightened failure probability, emphasizing the risks associated with excessive debt. Furthermore, RTAS and ROA, proxies for profitability relative to firm size, indicated that rising profits corresponded with reduced bankruptcy susceptibility, aligning with economic intuition.

These empirical findings affirm the validity of our applied technique, offering a practical solution for detecting business failure propensity while illuminating the algorithm's latent content. Importantly, our model not only classifies companies but also explains the rationale behind each classification at the local level. This allows for a detailed

understanding of how individual variables impact failure propensity, highlighting the individual weight and direction of each predictor for every observation.

From an ethical standpoint, our models are regarded as autonomous entities, rooted in prior programming. Consequently, ethical responsibility rests straight with individuals employing these technologies, rather than the models themselves. Given the model's robust performance and ease of interpretability, it holds potential value for diverse stakeholders seeking to identify and address potential failure situations, facilitating informed decision-making.

The use of break-down plots as a tool of explainable AI (XAI) at a local level, helps to expand on the importance of transparency and interpretability in ensuring fairness, accountability, and avoiding discriminatory practices. To avoid potential biases in the results of the algorithms, we encourage the need for human oversight in the decision-making process.

This study, while comprehensive in its exploration of XGBoost for business failure prediction, acknowledges certain limitations. The data used for analysis is specific to Spanish firms, potentially limiting generalizability to other contexts. Additionally, the focus on financial ratios might not encompass all factors contributing to business failure. Future research could broaden the scope by incorporating non-financial data such as corporate governance metrics or macroeconomic indicators. Furthermore, a comparative analysis of different machine learning algorithms could offer a more nuanced understanding of their strengths and weaknesses in this context.

Notably, the emphasis on interpretability in this study, facilitated by techniques like “break-down” plots, directly addresses ethical concerns surrounding the "black box" nature of ML models. By revealing the decision-making process and highlighting the contribution of individual features, we enhance transparency and foster trust in the model's predictions. This emphasis on interpretability not only strengthens the scientific rigor of our work but also contributes to the ethical and responsible use of AI in financial decision-making.

Future research should continue to prioritize interpretability and ethical considerations alongside predictive accuracy. This could involve using new explainable artificial intelligence techniques, and promoting interdisciplinary collaboration between

technical experts, ethicists, and policymakers. By proactively addressing these concerns, we can ensure that AI is utilized in a manner that benefits businesses, individuals, and society.

Interpretability in machine learning is an area of ongoing research and development. This study contributes to this effort by providing a practical example of how an XGBoost model can be interpreted and explained in the context of business failure prediction. We hope that this work will encourage further exploration and innovation in the field of interpretable machine learning for financial applications.

REFERENCES

Altman, E. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance 23(4), 589–609. http://dx.doi.org/10.1111/j.1540- 6261.1968.tb00843.

Barboza, F., Kimura, H., & Altman, E. I. (2017). Machine learning models and bankruptcy prediction. Expert Systems with Applications, 83, 405-417. https://doi.org/10.1016/j.eswa.2017.04.006

Beaver, W.H. (1966). Financial Ratios as Predictors of Failure, Journal of Accounting Research, 4, 71-111. https://doi.org/10.2307/2490171

Ben Braiek, H., & Khomh, F. (2024). Machine Learning Robustness: A Primer [arXiv eprint arXiv:2404.00897]. https://doi.org/10.48550/arXiv.2404.00897

Ben Jabeur, S., Stef, N., & Carmona, P. (2023). Bankruptcy prediction using the XGBoost algorithm and variable importance feature engineering. Computational Economics, 61(2), 715-741. https://doi.org/10.1007/s10614-021-10227-1

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. Computational Economics, 57(1), 203-216. https://doi.org/10.1007/s10614-020-10042-0

Candel, A & Ledell, E. (2018). Deep Learning with H2O. Retrieved June, 2024, from https://h2orelease.s3.amazonaws.com/h2o/rel-wheeler/4/docs-website/h2o-docs/booklets/DeepLearningBooklet.pdf

Carmona, P., Climent, F., & Momparler, A. (2019). Predicting failure in the US banking sector: An extreme gradient boosting approach. International Review of Economics & Finance, 61, 304-323. https://doi.org/10.1016/j.iref.2018.03.008

Carmona, P., Dwekat, A., Mardawi, (2022), Z.: No more black boxes! Explaining the predictions of a machine learning XGBoost classifier algorithm in business failure. Research in International Business and Finance, 61. https://doi.org/10.1016/j.ribaf.2022.101649.

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. Electronics, *8*(8), 832. https://doi.org/10.3390/electronics8080832

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SUGDD International Conference on Knowledge Discovery and Data Mining, 785-794, https://dl.acm.org/doi/10.1145/2939672.2939785

Cho, H., She, J., De Marchi, D., El-Zaatari, H., Barnes, E. L., Kahkoska, A. R., Kosorok, M. R., & Virkud, A. V. (2024). Machine learning and health science research: Tutorial. Journal of Medical Internet Research, 26(e50890), https://doi.org/10.2196/50890

Climent, F., Momparler, A., & Carmona, P. (2019). Anticipating bank distress in the Eurozone: An Extreme Gradient Boosting approach. Journal of Business Research, 101, 885-896. https://doi.org/10.1016/j.jbusres.2018.11.015

Dettling, M. y Bühlmann, P. (2003). Boosting for tumor classification with gene expression data, Bioinformatics, 19(9), 1061-1069. https://doi.org/10.1093/bioinformatics/btf867.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Du Jardin, P. (2010). Predicting bankruptcy using neural networks and other classification methods: The influence of variable selection techniques on model accuracy. *Neurocomputing*, *73*(10-12), 2047-2060. https://doi.org/10.1016/j.neucom.2009.11.034

Du Jardin, P. (2015). Dynamics of firm financial evolution and bankruptcy prediction. Expert Systems with Applications, 75, 25-43. https://doi.org/10.1016/j.eswa.2017.01.016

Du Jardin, P. (2016). A two-stage classification technique for bankruptcy prediction. European Journal of Operational Research, *254*(1), 236-252. https://doi.org/10.1016/j.ejor.2016.03.008

Du Jardin, P. (2017). Dynamics of firm financial evolution and bankruptcy prediction. Expert Systems with Applications, 75, 25-43. https://doi.org/10.1016/j.eswa.2017.01.016

Eling, M., & Jia, R. (2018). Business failure, efficiency and volatility: Evidence from the European insurance industry. International Review of Financial Analysis, 59, 58-76. https://doi.org/10.1016/j.irfa.2018.07.007

Erdogan, B. E. (2013). Prediction of bankruptcy using support vector machines: An application to bank bankruptcy. Journal of Statistical Computation and Simulation, 83(8). https://doi.org/10.1080/00949655.2012.666550

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

Friedman, J. (2002). Stochastic gradient boosting. Computational Statistics and Data Analysis, 38(4), 367-378. https://doi.org/10.1016/S0167-9473(01)00065-2

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). The Annals of Statistics, 28(2), 337-407. https://doi.org/10.1214/aos/1016218223

Frydman, H., Altman, E. I., & Kao, D. L. (1985). Introducing recursive partitioning for financial classification: the case of financial distress. The journal of finance, 40(1), 269-291. https://doi.org/10.1111/j.1540-6261.1985.tb04949.x

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An aproach to evaluating interpretability of machine learning. In 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA) (pp. 80-89). https://doi.org/10.1109/DSAA.2018.00018

Hall, P., & Gill, N. (2019). An introduction to machine learning interpretability (2nd Edition ed.). Sebastopol, CA. USA: O'Reilly Media, Incorporated.

Heaton, J., Goodfellow, I., Bengio, Y., & Courville, A. (2018). Deep learning. Genetic Programming and Evolvable Machines. Nature 19(1-2), 305-307. https://doi.org/10.1007/s10710-017-9314-z

Hosaka, T. (2019). Bankruptcy prediction using imaged financial ratios and convolutional neural networks. Expert Systems with Applications, 117. https://doi.org/10.1016/j.eswa.2018.09.039

Jabeur, S. B., Gharib, C., Mefteh-Wali, S., & Arfi, W. B. (2021). CatBoost model and artificial intelligence techniques for corporate failure prediction. Technological Forecasting and Social Change, 166. https://doi.org/10.1016/j.techfore.2021.120658

Jones, S. (2017). Corporate bankruptcy prediction: a high dimensional analysis. Review of Accounting Studies, 22(3), 1366-1422. https://doi.org/10.1007/s11142-017-9407-1

Kharkar, D. (2022). Unravelling the power of XGBoost: Boosting performance with extreme gradient boosting. Medium. Retrieved May , 2024, from https://medium.com/@dishantkharkar9/unravelling-the-power-of-xgboost-boosting-performance-with-extreme-gradient-boosting-302e1c00e555

Khoja, L., Chipulu, M., & Jayasekera, R. (2019). Analysis of financial distress cross countries: Using macroeconomic, industrial indicators and accounting data. International Review of Financial Analysis, 66. https://doi.org/10.1016/j.irfa.2019.101379

Le, H. H., & Viviani, J. L. (2018). Predicting bank failure: An improvement by implementing a machine-learning approach to classical financial ratios. Research in International Business and Finance, 44. https://doi.org/10.1016/j.ribaf.2017.07.104

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444. https://doi.org/10.1038/nature14539

LeDell, E., Gill, N., Aiello S. & Maloava, M. (2022). R Interface for the 'H2O' scalable machine learning platform. R package version 3.32.1.3. https://CRAN.R-project.org/package=h2o.

Lee, S., & Choi, W. S. (2013). A multi-industry bankruptcy prediction model using back-propagation neural network and multivariate discriminant analysis. Expert Systems with Applications, 40(8), 2941-2946. https://doi.org/10.1016/j.eswa.2012.12.009

Liang, D., Lu, C. C., Tsai, C. F., & Shih, G. A. (2016). Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. European Journal of Operational Research, 252(2), 561-572. https://doi.org/10.1016/j.ejor.2016.01.012

Merler, S., Furlanello, C., Larcher, B. & Sboner, A. (2001). Tuning Cost-sensitive Boosting and its Application to Melanoma Diagnosi. In International Workshop on Multiple Classifier Systems MCS2001, V. 2096 of lncs (pp. 32-42), Springer.

Momparler, A., Carmona, P., & Climent, F. (2016). La Predicción Del Fracaso Bancario Con La Metodología "Boosting Classification Tree". Revista Espanola de Financiacion y Contabilidad, 45(1), 63-91. https://doi.org/10.1080/02102412.2015.1118903

Momparler, A., Carmona, P., & Climent, F. (2020). Revisiting bank failure in the United States: a fuzzy-set analysis. Economic Research-Ekonomska Istrazivanja, 33(1), 3017-3033. https://doi.org/10.1080/1331677X.2019.1689838

Monasterio Astobiza, A. (2017). Ética algorítmica: Implicaciones éticas de una sociedad cada vez más gobernada por algoritmos. Dilemata, (24), 185–217. Recuperado a partir de https://www.dilemata.net/revista/index.php/dilemata/article/view/412000107

Mousavi, M. M., Ouenniche, J., & Xu, B. (2015). Performance evaluation of bankruptcy prediction models: An orientation-free super-efficiency DEA-based framework. International Review of Financial Analysis, 42, 64-75. https://doi.org/10.1016/j.irfa.2015.01.006

Mselmi, N., Lahiani, A., & Hamza, T. (2017). Financial distress prediction: The case of French small and medium-sized firms. International Review of Financial Analysis, 50, 67-80. https:// doi.org/10.1016/j.irfa.2017.02.004

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in neurorobotics, 7, 21. https://doi.org/10.3389/fnbot.2013.00021

Ohlson, J. (1980). Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research, 18(1), 109-131. https://doi.org/10.2307/2490395

ORBIS (2021). Base de datos de Bureau Van Dijk. A Moody's Analytics Company. https://authenticate.bvdep.com/rediris

Oyewo, B., Ajibola, O., & Ajape, M. (2020). Characteristics of consulting firms associated with the diffusion of big data analytics. Journal of Asian Business and Economic Studies, ahead-of-print(ahead-of-print). https://doi.org/10.1108/jabes-03-2020-0018

Perboli, G., & Arabnezhad, E. (2021). A Machine Learning-based DSS for mid and long-term company crisis prediction. Expert Systems with Applications, 174. https://doi.org/10.1016/j.eswa.2021.114758

Pozuelo, J., Martínez, J. & Carmona, P. (2018). Análisis de la utilidad del algoritmo Gradient Boosting Machine (GBM) en la predicción del fracaso empresarial. Spanish Journal of Finance and Accounting / Revista Española de Financiación y Contabilidad, 47(4), 507-532. https://doi.org/10.1080/02102412.2018.1442039.

Rapanyane, M. B., & Sethole, F. R. (2020). The rise of artificial intelligence and robots in the 4th Industrial Revolution: implications for future South African job creation. Contemporary Social Science, 15(4), 489-501. 10.1080/21582041.2020.1806346

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing. http://www.R-project.org/

Robnik-Šikonja, M. and Kononenko, I. (2008.) Explaining Classifications for Individual Instances. IEEE Transactions on Knowledge and Data Engineering, 20(5), 589–600. https://doi.org/10.1109/tkde.2007.190734

Romero, M., Carmona, P., & Pozuelo, J. (2021). Utilidad del Deep Learning en la predicción del fracaso empresarial en el ámbito europeo. Revista De Métodos Cuantitativos Para La Economía y la Empresa, 32, 392–414. https://doi.org/10.46661/revmetodoscuanteconempresa.5172

Santhanam, G. R., Holland, B., Kothari, S., & Ranade, N. (2017). Human-on-the-loop automation for detecting software side-channel vulnerabilities. In Information Systems Security: 13th International Conference, *ICISS 2017*, Mumbai, India (209-230). Springer International Publishing.

Sanz, J.A.. Galar, M., Bustince, H., Marco-Detchart, C., Gradín, C. & Belzunegui, T (2015). Predicción de la supervivencia de pacientes traumatizados graves utilizando EUSBoost. Actas de la XVI Conferencia de la Asociación Española para la Inteligencia Artificial. CAEPIA. Albacete.

Sigrist, F., & Leuenberger, N. (2023). Machine learning for corporate default risk: Multi-period prediction, frailty correlation, loan portfolios, and tail probabilities. European Journal of Operational Research, 305(3), 1390-1406. https://doi.org/10.1016/j.ejor.2022.06.035

Smith, M., & Alvarez, F. (2021). A machine learning research template for binary classification problems and shapley values integration. Software Impacts, 8. https://doi.org/10.1016/j.simpa.2021.100074

Staniak, M., & Biecek, P. (2019). Explanations of Model Predictions with live and breakDown Packages. The R Journal, 10(2), 395. https://doi.org/10.32614/RJ-2018-072

Stef, N. (2021). Institutions and corporate financial distress in Central and Eastern Europe. European Journal of Law and Economics, *52*(1), 57-87. https://doi.org/10.1007/s10657-021-09702-9

Syam, N., & Sharma, A. (2018). Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelligence in sales research and practice. Industrial Marketing Management, 69, 135-146. https://doi.org/10.1016/j.indmarman.2017.12.019

Tascón, M.T. and Castaño, F.J. (2012). Variables y Modelos para la Identificación y Predicción del Fracaso Empresarial: Revisión de la Investigación Empírica Reciente, Revista de Contabilidad, 15(1), 7-58. https://doi.org/10.1016/S1138-4891(12)70037-7

Tian, S., & Yu, Y. (2017). Financial ratios and bankruptcy predictions: An international evidence. International Review of Economics and Finance, 51. https://doi.org/10.1016/j.iref.2017.07.025

Tsai, C. F., & Cheng, K. C. (2012). Simple instance selection for bankruptcy prediction. Knowledge-Based Systems, 27, 333-342. https://doi.org/10.1016/j.knosys.2011.09.017

Vega García, M., & Aznarte, J. L. (2020). Shapley additive explanations for NO2 forecasting. Ecological Informatics, 56, 101039. https://doi.org/10.1016/j.ecoinf.2019.101039

Wilson, G. I. & Sharda, R. (1994). Bankruptcy prediction using neural network. Decision Support Systems, 11, 545-557. https://doi.org/10.1016/0167-9236(94)90024-8

Xia, Y., Liu, C., Li, Y. Y., & Liu, N. (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Systems with Applications, 78, 225-241. https://doi.org/10.1016/j.eswa.2017.02.017

Zhou, L., Si, Y. W., & Fujita, H. (2017). Predicting the listing statuses of Chinese-listed companies using decision trees combined with an improved filter feature selection method. Knowledge-Based Systems, 128, 93-101. https://doi.org/10.1016/j.knosys.2017.05.003

Zięba, M., Tomczak, S. K., & Tomczak, J. M. (2016). Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. Expert Systems with Applications, 58, 93-101. https://doi.org/10.1016/j.eswa.2016.04.001

Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. Journal of Accounting Review, 22, 59-82. https://doi.org/10.2307/2490859