Bankruptcy Prediction using the XGBoost Algorithm and Variable Importance Feature Engineering

Sami Ben Jabeur · Nicolae Stef · Pedro Carmona

Abstract
The emergence of big data, information technology, and social media provides an enormous amount of information about firms' current financial health. When facing this abundance of data, decision makers must identify the crucial information to build upon an effective and operative prediction model with a high quality of the estimated output. The feature selection technique can be used to select significant variables without lowering the quality of performance classification. In addition, one of the main goals of bankruptcy prediction is to identify the model specification with the strongest explanatory power. Building on this premise, an improved XGBoost algorithm based on feature importance selection (FS-XGBoost) is proposed. FS-XGBoost is compared with seven machine learning algorithms based on three well-known feature selection methods that are frequently used in bankruptcy prediction: stepwise discriminant analysis, stepwise logistic regression, and partial least squares discriminant analysis (PLS-DA). Our experimental results confirm that FS-XGBoost provides more accurate predictions, outperforming traditional feature selection methods.

Keywords Corporate failure · XGBoost · Machine learning · Bankruptcy

1 Introduction

Assessing the risk of business failure has long been a major concern for researchers, bankruptcy practitioners, accountants and banks. Financial failure affects company survival and imposes high costs on bankers and other creditors, who may only partially recover their investments (Wruck, 1990). In general, failure refers to situations that lead to firm's bankruptcy following payment default. Everett and Watson (1998) showed that the concept of failure is associated with three types of risk: risk related to the national economy, risk related to the firm's industry, and risk that is unique to the business itself. Moreover, the process of failure may vary. As noted by Laitinen et al. (2014), firms that fail may decline quickly, gradually collapse, or operate poorly for a long period. When business failure is imminent, debtors (firms) can find shelter in the bankruptcy system. For instance, reorganization can encourage firm survival despite financial failure (Blazy & Stef, 2020; Stef, 2018, 2021).

Unsurprisingly, corporate failure prediction is a serious problem in risk management, and it is treated carefully by banks and other financial institutions. Consequently, financial organizations must develop effective warning systems to predict bankruptcy (Liang et al., 2015). The emergence of big data, information technology, and social media provides an enormous amount of information about firms' current financial health. When faced with this abundance of data, decision makers must identify the crucial information to build an effective and operative prediction model without sacrificing the quality of the estimated output. Achieving such a goal is enabled by feature selection, which is an important data preprocessing stage in machine learning models (Liang et al., 2015).

In this paper, an improved Extreme Gradient Boosting (XGBoost) algorithm based on feature importance selection (FS-XGBoost) is proposed to predict corporate failure. The primary contributions of this paper are twofold. First, we combine a feature importance strategy with the XGBoost algorithm. Our estimates suggest that FS-XGBoost can provide enhanced accuracy and suitability in financial distress prediction. Second, we compare FS-XGBoost with seven machine learning models. We thus identify which classification methods are most sensitive to the feature selection technique. Therefore, our study offers banks and financial institutions guidelines in their search for suitable bankruptcy prediction models.

This paper is structured as follows. Section 2 presents the literature review. Section 3 describes the methods used in our study. Section 4 presents our sample and the variables used in the empirical analysis. Section 5 summarizes and discusses the experimental results. Section 6 concludes.

2 The Landscape of Bankruptcy Prediction Techniques

Early warning systems are tools that enable financial institutions to promote stability more efficiently in a financial system. The foundations were laid by Altman (1968) and Bardos (1998), who investigated the power of financial ratios in forecasting corporate failure using discriminant analysis. However, numerous studies have aimed at obtaining more accurate predictive models. For example, the accuracy of discriminant analysis has been compared with that of other classes of models (i.e., probabilistic models, Multi-Layer Perceptron, and various machine learning algorithms).

Certain researchers have concluded that intelligence modeling techniques provide the best performance when compared with discriminant analysis and logistic regression models (Lee et al., 2005; Chen et al., 2009; du Jardin, 2017; Kim & Kang, 2010; Tsakonas et al., 2006). For example, du Jardin (2010) found that neural networks provide better results than the discriminant analysis and logistic regression in relation to identifying potentially failing firms. Similarly, Kim et al. (2016) examined the effectiveness of clustering techniques and genetic algorithms based on artificial neural networks. Their results show that a hybrid model is the most suitable technique for imbalanced data. According to Lee and Choi (2013), neural network outperforms multivariate discriminant analysis in corporate bankruptcy in the case of Korean firms. Moreover, du Jardin and Séverin (2012) examined the power of the Kohonen map model, which mainly involves a set of neurons structured on a square grid. Their findings again confirmed that the accuracy of a map model approach is greater than that discriminant analysis, logistic regression and Cox's model.

Several machine learning techniques have been successfully applied in bankruptcy prediction. Shin et al. (2005) used support vector machines (SVMs) in financial distress prediction, showing that the SVM approach outperforms the back-propagation neural network in terms of accuracy. Considering the diverse perspective of estimation approaches, Tsai and Cheng (2012) compared the predictive power of support vector machines, discriminant analysis, logit regression, and neural networks in tackling bankruptcy prediction problem. They found that, on average, SVMs offer the most efficient technique in terms of accuracy for predicting corporate bankruptcy. Zhao et al. (2017) used a Kernel Extreme Learning Machine (KELM) to discriminate bankrupt companies against non-bankrupt companies. Their results showed that KELM performs better than Support Vector Machines, Extreme Learning Machines and Random Forest. Among statistical techniques, decision trees are also widely used in corporate failure prediction (Geng et al., 2015; Olson et al., 2012; Ravi Kumar & Ravi, 2007; Zhou et al., 2017). Nevertheless, the range of techniques is actually much wider. Other machine learning methods have been employed including boosting, bagging, and random forest models (Barboza et al., 2017; Choi et al., 2018; Jabeur et al., 2020; Kim & Kang, 2010; Wang et al., 2014). In this frame, it should be pointed out the paper of Zhou and Lai (2017) who applied AdaBoost on corporate bankruptcy prediction with missing values revealing that it performs better than other benchmark models.

One of the most innovative forecasting approaches was developed by Xia et al. (2017). They proposed a novel boosted tree for a credit scoring model based on Extreme Gradient Boosting (XGBoost). Climent et al. (2019) used XGBoost to identify the financial indicators that can be used to forecast bank failure in the Eurozone banking sector. According to Climent et al. (2019), this technique performs better than logistic regression, random forest, Multi-Layer Perceptron, and Light Gradient Boosting (LightGBM). However, another important prediction tool based on recent advances in machine learning techniques relates to deep learning models. Using a sample of 11,827 U.S. public companies, Mai et al. (2019) showed that deep learning models perform better than statistical models when forecasting financial distress. By considering the periods 1, 2 and 3 years prior to the insolvency, Fernández-Gámez et al. (2016) applied a deep learning model using financial and non-financial variables on a sample of Spanish hotels that filed for bankruptcy between 2005 and 2012. In addition, Mai et al. (2019) applied a convolutional neural network to predict corporate bankruptcy, confirming that this method also outperforms conventional methods.

More recently, bankruptcy prediction studies have started to focus on feature selection techniques (Son et al., 2019). Selected features could help researchers overcome shortcomings such as multicollinearity (Jabeur, 2017; Yang et al., 2011), optimal features (Zhang et al., 2019a, b), and dimensionality (Bolón-Canedo & Alonso-Betanzos, 2019). Using partial least squares (PLS), Yang et al. (2011) found that PLS-based feature selection can successfully identify complex nonlinearity and correlations among accounting ratios. Brezigar-Masten and Masten (2012) proposed nonparametric regression and the classification tree (CART) method for feature selection problems. Seemingly, the nonparametric CART method has shown to yield higher classification accuracy compared to stepwise logit. Furthermore, Liang et al. (2014) investigated the effects of feature selection on bankruptcy prediction using discriminant analysis, t-tests, and logistic regression. No optimal combination of the feature selection technique and the classification approach was confirmed by their study. Nevertheless, Lin et al. (2014) proposed the wrapper method to reduce the number of financial ratios in financial distress prediction. Their analysis showed that the wrapper method significantly outperforms conventional feature selection approaches. Similarly, Liang et al. (2014) examined the effect of filter- and wrapper-based feature selection methods on bankruptcy prediction. However, they concluded that feature selection does not always improve the prediction performance of classification models.

3 Modeling Methods

In this section, we describe seven classification techniques used to distinguish between failed and nonfailed companies. These classification techniques are based on four feature selection techniques: stepwise logistic regression, stepwise discriminant analysis, partial least squares discriminant analysis, and XGBoost. We aimed to predict bankruptcy likelihood for French non-listed firms. Figure 1 provides the conceptual framework, including the dataset, training and testing sample, feature extraction and selection, model-performance comparison, and evaluation of feature importance. First, we constructed a data set using DIANE database of Bureau Van Dijk. Second, we used chained equations for imputation. Third, four feature-selection methods were adopted to select the optimal feature**.** Finally, we applied seven machine learning models and we evaluated the importance of each feature.

3.1 Discriminant Analysis

Discriminant analysis discriminates between observations based on their individual characteristics. This technique is used to classify and predict qualitative variables. The advantage of this method over univariate analysis is that it enables discrimination

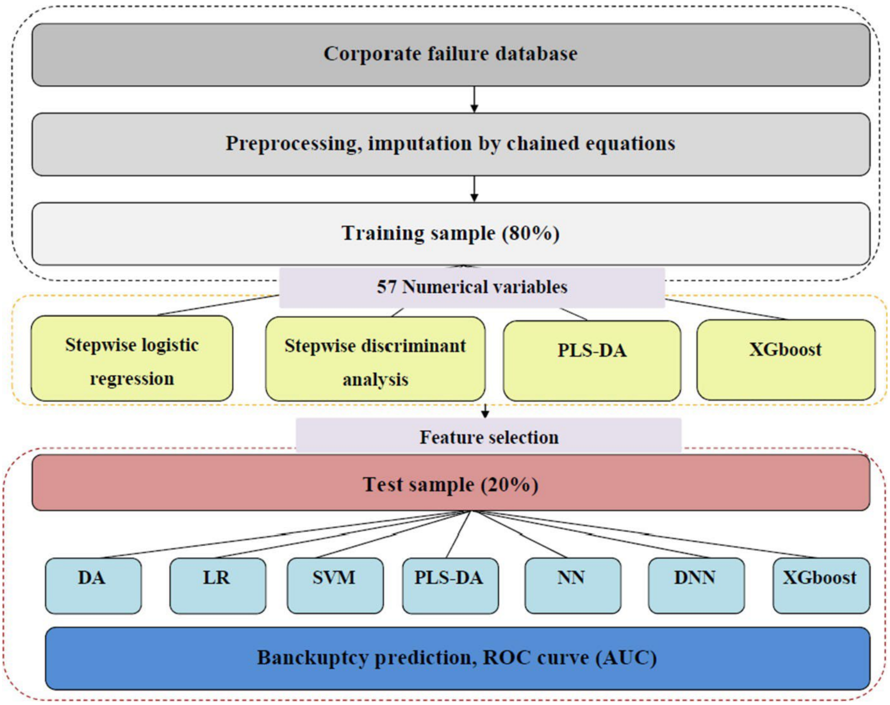

Fig. 1 Flowchart of the experimental procedure. The original data set was divided into two subsets: a training set comprising 80% of the data, and a test data set comprising 20% of the data. AUC estimated by type of feature selection method

between the possible outcomes by simultaneously using all indicators. According to Altman (1968), this method consists of finding the weighted average of several ratios calculated for each company (i.e., the discriminant function) that best distinguishes between failing and healthy firms. The function estimates a Z-score for the firms. This score can be expressed as follows:

$$Z = \beta_0 + \sum_{i=1}^{N} \beta_i X_i$$

where β_i represents the discriminant weights, X_i represent the accounting ratios, N the number of features and β_0 is a constant. Discriminant analysis requires the distinction between failing and healthy firms to be linearly separable, treating the ratios as if they were independent (Barboza et al., 2017; du Jardin, 2016).

3.2 Logistic Regression

Other alternatives for corporate failure prediction based on logistic regression have been applied by Ohlson (1980). Logistic regression models link a qualitative dependent variable to a set of explanatory variables. The qualitative dependent variable is used to define membership of the data to a certain category (failed company vs. healthy company). The probability function of failure is calculated as follows:

$$Z = \frac{1}{1 + e^{-\sum_{i=1}^{N} \beta_i X_i + \beta_0}}$$

where X_i captures the explanatory financial ratios, and β_i are the coefficients of the variables. In the case of univariate dichotomous models, several estimation methods are available. The most widely used technique is maximum likelihood. As with conventional linear regression, the presence of correlations between predictors leads to anomalies in the estimation (Jabeur, 2017).

3.3 Support Vector Machines

Support vector machines (SVMs) represent a set of related supervised learning methods that are widely used for classification problems. The original algorithm was proposed by Cortes & Vapnik, (1995). Support vector machines seek to identify the optimal separating hyperplane among failed and nonfailed firms by maximizing the margin between them. According to Bao et al. (2019), different algorithms in SVMs can be applied to draw input variables into a high-dimensional feature space. In addition, coefficients can be expressed and estimated using the maximum likelihood method. In the case of bankruptcy prediction, the probability of failure can be written as follows:

$$P[Y_i = 1/X_i] = \frac{1}{1 + \exp[-f(X_i)]}$$

where X_i are explanatory variables, Y_i are dependent variables, and f(x) is the polynomial function. Nevertheless, SVMs have some shortcomings such difficulties identifying the correct parameter values, difficulties processing large data sets, and the noisiness of classes.

3.4 Partial Least Squares Discriminant Analysis

Partial least squares (PLS) regression has been widely used in various disciplines such as marketing, business, economics, and finance (Becker & Ismail, 2016; Bellini et al., 2017; Sghaier et al., 2018). This data analysis method is specifically designed to study the relationships between a set of response variables, Y, and a set of explanatory variables, X, when multicollinearity is severe. In PLS regression, principal component analysis is performed for each set of variables X and Y, under the constraint that the components of X are strongly correlated with those of Y (Bastien et al., 2005). The Z-score of this technique is given as follows:

$$Z = \sum_{h=1}^{M} C_h T_h$$

where C_h are the estimated coefficients, M is the number of components and T_h are the PLS components. PLS regression is performed in the first stage to reduce data dimensionality and extract PLS components. In the second stage, discriminant analysis is used to estimate the Z-score.

3.5 Neural Networks

Neural networks (NN) are a class of powerful universal tools that can be used for prediction and classification. They are already applied in various fields, especially medical diagnosis (Alzubi et al., 2019; Daoud et al., 2019), credit card fraud (Ander et al., 2018; Kim et al., 2019; Leong et al., 2013), and corporate failure (du Jardin, 2010; Hernandez Tinoco & Wilson, 2013; Ravisankar & Ravi, 2010). Multilayer perceptron models are the most commonly used models. Several processing layers allow them to make nonlinear associations between inputs and output. The output Z-score is given by:

$$Z = \sum_{i=1}^{N} f\left(\sum_{j=1}^{P} w_{ij} X_i + \beta_j\right) w_{jk} + \beta_k$$

where W_{ij} and W_{jk} correspond to the weights of the hidden layer and the output layer, f is the activation function, N represents the number of variables, P the number of hidden neurons, X_i are the input variables, β_j is the bias value of the hidden neurons, and β_k is the bias value of the output.

3.6 Deep Neural Networks

Deep learning is a new branch of machine learning that has been applied to various classification problems in areas such as medicine (Hu et al., 2018; Kumar et al., 2019), speech recognition (Boloukian & Safi-esfahani, 2019; Qawaqneh et al., 2017), and finance (Chatzis et al., 2018; Kraus & Feuerriegel, 2017). A deep neural network (DNN) is a new alternative to artificial neural networks. It works with multiple hidden layers of units, which are present between the input and output layers (Dixon et al., 2015). Hinton (2006) used a DNN with a widely Restricted Boltzmann Machine (RBM) architecture. RBMs are used to create stochastic models of neural networks that can compute the probability of failure according to the input ratios. The probability of failure is defined as:

$$P[X, \mu] = \frac{\exp(-E(X, \mu))}{Z}$$

where Z is the partition function, μ is the hidden vector, and E is the energy function of the RBM. According to Hinton (2006), the contrastive divergence algorithm is an effective way to estimate the parameters of RBM.

3.7 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) models have recently attracted considerable attention in credit scoring and bankruptcy prediction (Carmona et al., 2019; Chang et al., 2018; Climent et al., 2019; Son et al., 2019). According to Chen and Guestrin (2016), XGBoost combines regression trees and gradient boosting. At each tree of the training process, the residual of a base classifier is employed in the next classifier to improve the objective function. The XGBoost algorithm reduces the complexity of modeling and prevents the problems associated with overfitting. Finally, the combination of all trees provides the final target. The prediction output is given as follow:

$$Z = G(X_i) = \sum_{j=1}^{K} g_j(X_i)$$

where X_i are the financial ratios, and $g_k(X_i)$ is the output function of each tree. Following Zhang, Qiu, et al. (2019), b) and Son et al. (2019), we describe the procedures of XGBoost in Algorithm 1.

Compared to more recent algorithms, such as CatBoost or LightGBM, XGBoost cannot handle categorical features (Jabeur et al., 2021) as it only runs numerical values similar to Random Forest. Therefore one has to execute several encodings, such as label encoding, mean encoding or one-hot encoding before bringing categorical data to XGBoost.

Algorithm 1: XGBoost algorithm

Input: Explanatory feature matrix: X; target attribute vector: Y; loss function: $l(y, \hat{y})$; base learner: $g(X, \mu)$; number of subtrees: K

Output: Algorithm

Setup 1: for t = 1:K do

Setup 2: Initialize $G_0(X_i) = \text{argmin}_\rho = \sum_{i=1}^{N} l(y_i, \rho)$

Setup 3: Compute $\nabla G_t(X)$

Setup 4: Run a new learner function $g(X, \mu)$

Setup 5: Predict the best gradient descent stage size $\rho_k = argmin\_\rho \sum_{i=1}^{N} l(y_i, \hat{G}_{k-1}(X_i) + \rho g(X_i, \mu_i)$

Setup 6: Output the prediction probability: $\hat{G}_k = \hat{G}_{k-1} + \rho_k g_k(X, \mu)$

3.8 FS-XGBoost Algorithm

In certain situations, the original features were redundant and noisy, which could have a detrimental effect on the model training phase. Therefore, a robust feature extraction technique needs to be employed for effective classification work (Yu et al., 2019). According to Jones (2017), the predictor variables usually have multiple impacts on the outcome's domain of mathematical applications. It is also helpful to identify their variable importance that displays the relative contribution of each input variable to the overall output of the model. For XGBoost, the relative importance of the features can be measured by different approaches, namely split weight, mean gain, etc. Following

XGBoost fitting, the feature rankings of weight-based and gain-based importance can be obtained. According to Shi et al. (2019), the optimal feature subset can be selected between output learning and model simplicity (i.e., fewer features) depending on the trade-off. It is possible to delete redundant and unnecessary variables without damaging the accuracy. Feature selection provides improved interpretability, simplified modelling, shorter learning time, and better generalizations (García, Luengo, et al., 2016; García, Ramírez-Gallego et al., 2016). In order to achieve this purpose, the FS-XGboost approach was used to minimize dimensionality. This algorithm will not only increase the efficiency of the classifier in terms of simplicity, but also strengthen the model generalization and interpretability. According to Occam's razor, or the law of parsimony, for any two models that explain the occurrence of a certain event, the simpler one is preferable. When XGBoost score is larger, the corresponding feature vector will be more important. In other words, all input functions shall be ordered by scores in descending order. We selected the top-ranked feature whose score was higher than k for the feature sets, as illustrated in Algorithm 2.

Algorithm 2: FS-XGBoost algorithm

1. Train XGBoost based on (X,Y);
2. Tune model (XGBoost) using all features (X,Y);
3. Rank features relative importance;
4. Remove less importance features;
5. Find the best feature subset;
6. Data set with key features (X, Y)';
7. Model training and hyper-parameters tuning based on (X, Y)';
8. Prediction performance evaluation.

4 Data and Variables

The construction of the forecasting models was based on a sample of small and medium sized French firms from different sectors. The sample was divided into two groups: healthy companies and failing companies. The DIANE database (Bureau Van Dijk) was used to randomly draw the sample. We drew the sample of failed firms from the list of firms that filed for bankruptcy in 2017. The DIANE database reports 552 firms that were subject to a liquidation procedure in that year. However, only 34.4% of those failed firms presented complete accounting data available for one year before failure (2016), two years before failure (2015), and three years before failure (2014). In addition, the DIANE database includes 84,872 firms that were not subject to a bankruptcy procedure in 2017. Among those firms, only 1660 firms reported complete accounting data over the period 2009–2017. The reporting of the financial balance sheets and the income statements on an annual basis implies that firm's activity was not previously suspended providing consistency to the inclusion of those firms in the group of healthy firms. The final sample following this rigorous selection process consisted of 1850 companies divided into two subsamples: 1660 healthy companies and 190 failing companies.

The sample is imbalance because the number of failed companies is much lower than the number of non-failed companies. An imbalance occurs when one or more classes have very low proportions in the training data as compared to the other classes. There are different strategies for overcoming class imbalances: most researchers attempt to solve this challenge with sampling methods to balance the class frequencies, either up-sampling or down-sampling the data. In our study, we have chosen the former method by means of simulating or imputing additional data points to improve balance across classes. Taking this approach eliminates the fundamental imbalance issue that affects model training.

The financial ratios were selected based on their recurrence in the literature (Altman, 1968; du Jardin, 2017; Liang et al., 2016; Mai et al., 2019; Ravi Kumar & Ravi, 2007) and their relevance to the financial analysis. Thus, we selected key ratios for most bankruptcy detection models. The selected ratios were solvency, liquidity, profitability, productivity, turnover, and financial structure. A set of 57 ratios and variables (R01 to R57) were selected (see Table 1). These ratios are commonly used in the literature, and they provided substantive information to enable analysis of the financial situation of French firms.

5 Results

5.1 Data Preprocessing

Data preprocessing is one of the most important steps when using machine learning techniques (Cordón et al., 2019; García, Luengo, et al., 2016; García, Ramírez-Gallego, et al., 2016; Krawczyk & Herrera, 2017). The main objective of data preprocessing is to clean and prepare the input data to ensure the data are correctly employed in the classification and regression problem. Most bankruptcy databases have missing data. Scholars have strong incentives to remove observations with missing values. To fill missing values, we used imputation by chained equations, employing the *mice* package (Buuren & Groothuis-Oudshoorn, 2010) in R software. We only used imputation for variables with less than 30% of missing data. The remaining variables were eliminated. We also eliminated observations with more than 25% of missing values. To avoid overfitting, the data set was divided into two subsets: a training set comprising 80% of the data, and a test data set comprising 20% of the data. In this study, performance metrics were provided from the test sample. The stages of the experimental procedure are shown in Fig. 1.

5.2 Feature Selection Results

We ran the seven machine learning models based on four feature selection methods: XGBoost, stepwise logistic regression, stepwise discriminant analysis, and partial least squares discriminant analysis.

Table 2 and Fig. 2 show the importance of the 10 most relevant ratios according to XGBoost. Based on the partial dependence plots in Fig. 3, XGBoost allows managers

Table 1 Set of features and definitions

Ratio	Definition
R02	Shareholder liquidity ratio
R03	Solvency ratio
R04	Gearing ratio
R05	Equity per employee
R06	Working capital per employee
R07	Total assets per employee
R08	Global performance as measured by Diane
R09	Sustainable return on investment
R10	Earnings before taxes (EBT)/total assets
R11	Inventory turnover
R12	Accounts receivable
R13	Accounts payable
R14	Total net assets turnover
R15	Staff costs/turnover
R16	Turnover per employee
R17	Annual average wage
R18	Net income per employee
R19	Earnings before interest and taxes/total assets
R20	Earnings before interest, taxes, depreciation, and amortization (EBITDA)/total assets
R21	EBITDA/number of employees
R22	Added value/fixed assets
R23	Added value/financial capital
R24	Added value/equity
R25	Investment rate
R26	Net profitability
R27	Return on net EQUITY
R28	Added value/turnover
R29	Inventory turnover
R30	Turnover/employees (thousand euro/employee)
R31	Financial interest rate
R32	Interest/turnover
R33	Leverage
R34	Repayment capacity
R35	Cash flow
R36	Working capital/turnover
R37	Current ratio
R38	Quick ratio

Table 1 (continued)

Ratio	Definition
R39	Total assets
R40	EBITDA
R41	EBT
R42	Working capital
R43	Export turnover
R44	Return on net assets
R45	Return on equity
R46	Export ratio
R47	Number of employees
R48	Unemployment rate of firm's department
R49	Dummy variable equal to 1 for industrial firms, 0 otherwise
R50	Dummy variable equal to 1 for trading firms, 0 otherwise
R51	Dummy variable equal to 1 for agricultural firms, 0 otherwise
R52	Dummy variable equal to 1 for firms operating in other sectors, 0 otherwise
R53	Dummy variable equal to 1 for limited liability firms, 0 otherwise
R54	Dummy variable equal to 1 for limited liability firms with sole owner, 0 otherwise
R55	Dummy variable equal to 1 for simplified joint stock firms, 0 otherwise
R56	Dummy variable equal to 1 for firms of different legal form, 0 otherwise
R57	Age of the firm

and financial institutions to determine the complex marginal dependence of each feature on the estimated target (Friedman, 2001; Son et al., 2019). Many machine learning algorithms, such as neural network, have been criticized because they cannot provide real dependencies of the prediction on the features (Jabeur et al., 2021), meaning that their interpretation is not always easy. Nevertheless, the partial dependency plot can demonstrate more complex marginal dependency of each feature on the predicted outcome of an advance machine learning model. According to Son et al. (2019) the partial dependence function is calculated as follows:

$$f_{x_s}(x_s) = E_c(\hat{f}(x_s, x_c))$$

where X_s are the variables of interest, X_c are the other variables used in the machine learning model, and f is the machine learning technique or the algorithm. Partial dependency plots functions by marginalizing the predicted output over the features that we are not involved in, so that the function reveals the marginal dependence of the accounting variables that we are involved in (Son et al., 2019).

For a 1-year horizon, there is a positive relationship between the number of employees (R47) and corporate failure. These results are consistent with those reported by Platt and Platt (1994), who identified a significant impact of employment on corporate failure rates and a negative relationship between bankruptcy and the solvency ratio

Table 2 Feature selection results

Selected ratios	Discriminant analysis			Logistic regression			PLS-DA			XGBoost		
	Stepwise method			Stepwise regression			Variable importance projection (VIP)			Variable importance		
	1 year (1)	2 years (2)	3 years (3)	1 year (4)	2 years (5)	3 years (6)	1 year (7)	2 years (8)	3 years (9)	1 year (10)	2 years (11)	3 years (12)
R01							1.128	1.287	1.303			
R03	0.781***	0.850***	0.855***	− 0.035***	− 0.019***	− 0.026***	2.918	3.151	3.717	0.117	0.168	0.129
R04					0.002***		1.539	1.216	1.773			
R05									1.714	0.034		
R07									1.889	0.045		
R08							1.789	1.728				0.043
R09	0.711***	0.789***	0.820***				2.209	1.258		0.032		
R10			0.828***			− 0.048***	1.798	1.196				
R12					0.003***							
R14							1.128	1.3892	1.910			0.027
R15	0.693***	0.780***								0.039	0.026	
R19							1.395	1.150	1.214			
R20					− 0.029***		1.443	1.241	1.327			
R21										0.038		
R24	0.691***	0.798***	0.851***		1.969***	1.624***	1.414	2.043	2.397		0.065	0.105
R26											0.023	0.054

Table 2 (continued)

Selected ratios	Discriminant analysis			Logistic regression			PLS-DA			XGBoost		
	Stepwise method			Stepwise regression			Variable importance projection (VIP)			Variable importance		
	1 year (1)	2 years (2)	3 years (3)	1 year (4)	2 years (5)	3 years (6)	1 year (7)	2 years (8)	3 years (9)	1 year (10)	2 years (11)	3 years (12)
R29	0.686***	0.778**										
R30												0.031
R34			0.817**			− 0.012**						
R35							1.595	1.539	1.473	0.056		
R37							1.191	1.447	1.393	0.042	0.062	
R38					− 1.017***		1.205	1.457	1.389	0.077	0.055	0.099
R39				− 0.0001***							0.028	0.054
R40	0.686***	0.778**		− 0.0007***								
R41											0.028	0.034
R42											0.026	0.047
R43	0.684***			− 0.0001								
R44					− 0.0006***							
R45		0.779***	0.824***			0.0001**			1.168			
R47	0.739***	0.831***		0.087***	0.027***		2.688	2.846		0.083	0.198	

Table 2 (continued)

Selected ratios	Discriminant analysis			Logistic regression			PLS-DA			XGBoost		
	Stepwise method			Stepwise regression			Variable importance projection (VIP)			Variable importance		
	1 year (1)	2 years (2)	3 years (3)	1 year (4)	2 years (5)	3 years (6)	1 year (7)	2 years (8)	3 years (9)	1 year (10)	2 years (11)	3 years (12)
R48	0.685***		0.816**									
R52	0.683***											
R53	0.683***											

Columns (1), (4), (7), and (10) use the 1-year lagged values, columns (2), (5), (8), and (11) the 2-year lagged values, and columns (3), (6), (9), and (12) the 3-year lagged financial ratios

*At the 10% level

**At the 5% level

***Denotes a significant coefficient at the 1% level

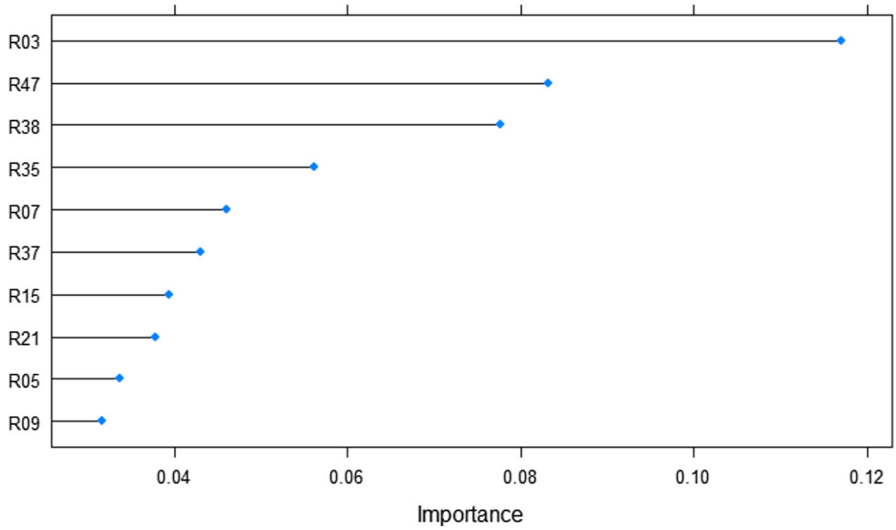

Fig. 2 The 10 most important features 1 year before failure

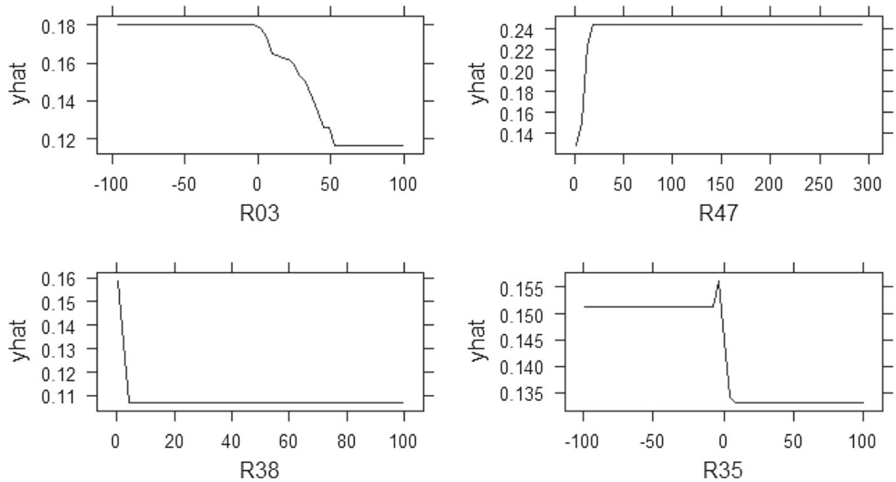

Fig. 3 Partial dependence plots for the four most influential variables 1 year before failure

(R03), cash flow (R35), and quick ratio (R38). These results are robust with respect to those of Stef and Jabeur (2018), who identified similar significant ratios for a sample of nonlisted French firms. The PLS-DA feature selection technique confirms the previous results, identifying 14 important variables based on the variable importance in projection (VIP) scores (Wold, 1985). VIP scores can be used to rank the most important ratios according to their explanatory power with respect to the target variable (VIP ≥ 1). Stepwise logistic regression and discriminant analysis selected fewer

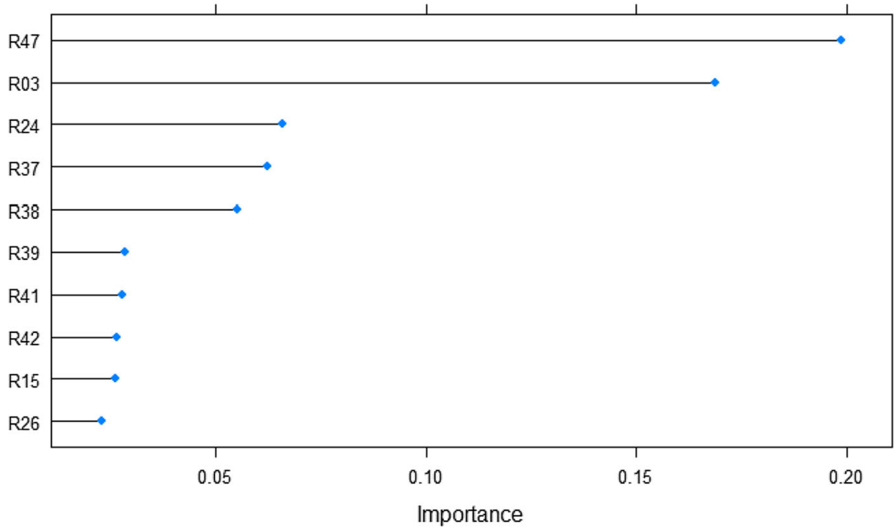

Fig. 4 The 10 most important features 2 years before failure

variables than PLS-DA and XGBoost because of the high correlations between ratios. These results support those reported by Yang et al. (2011) and Serrano-cinca and Gutiérrez-Nieto (2013), who argued that traditional models suffer from problems of multicollinearity, especially among financial explanatory variables. Both the logistic regression and the discriminant approach selected the solvency ratio (R03), EBITDA (R40), export turnover (R43), and number of employees (R47). The analysis indicates that having high solvency and profitability ratios reduces the probability of corporate failure.

Similarly, for two years before failure, XGBoost shows positive relationships between added value/equity (R24) and number of employees (R47) and corporate failure (Figs. 4, 5), and negative relationships between the solvency ratio and bankruptcy. The lower a company's solvency ratio (R03) is, the higher the likelihood that the company will default on its current obligations. In addition, PLS-DA shows that two years before failure, 15 ratios play a crucial role in discriminating between healthy and failing firms. The solvency ratio (R03) is ranked first, with a variance inflation factor (VIF) of 3.151. Solvency reflects a company's ability to meet debt repayment deadlines. Moreover, the cash flow ratio (R35) and current ratio (R37) indicate that healthy companies have a higher level of liquidity than companies in default. This finding may seem obvious because failing companies struggle to meet their financial commitments. Indeed, when solvency deteriorates, banks become more entrenched and have greater bargaining power over their debtors (firms). Banks' risk aversion encourages the adoption of measures that tighten credit conditions. Finally, lower liquidity renders a company unable to settle creditors' claims, thereby leading to a significant decrease in the company's profitability.

Analysis for three years before failure (Table 1; Fig. 6) shows that XGBoost selected 10 key ratios. Figure 7 provides the partial dependence plots for the four most important

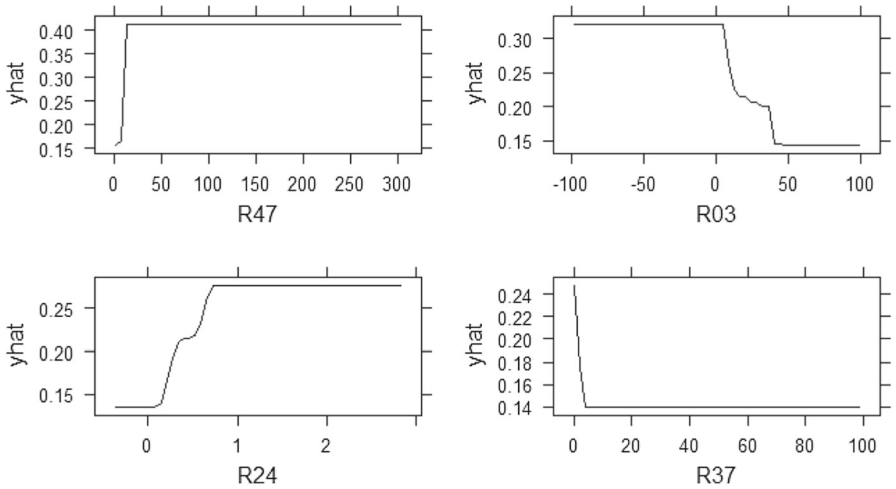

Fig. 5 Partial dependence plots for the four most influential variables 2 years before failure

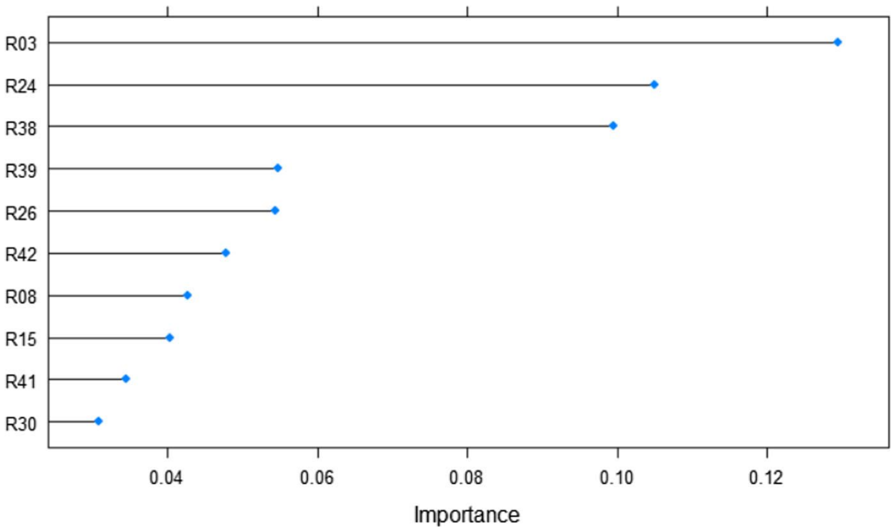

Fig. 6 The 10 most important features 3 years before failure

ratios. There is a positive relationship between financial distress and added value/equity (R24) and total assets (R39) and a negative relationship between bankruptcy and the solvency ratio (R03) and quick ratio (R38). More specifically, the results indicate that higher working capital levels are associated with a higher risk of failure. Conversely, the likelihood of corporate failure is lower when a company has more liquid assets (Climent et al., 2019). This finding is consistent with most of the findings reported in

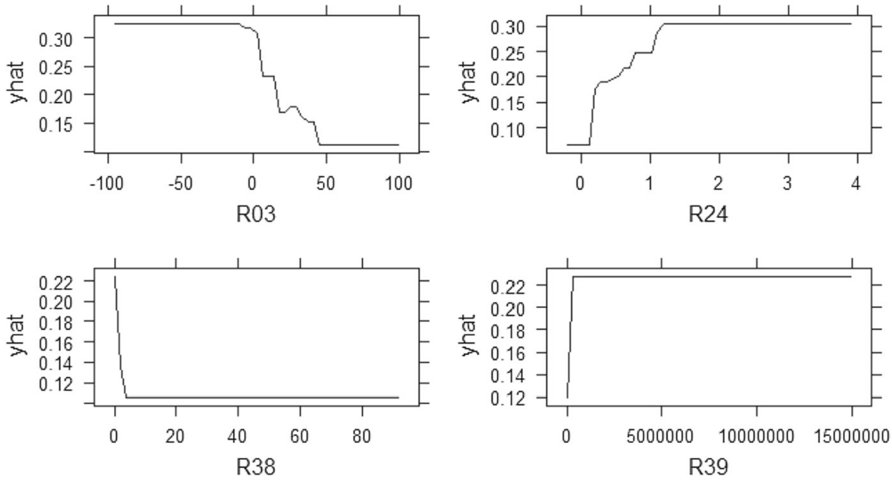

Fig. 7 Partial dependence plots for the four most influential variables 3 years before failure

the literature (e.g., Altman, 1968; Jones, 2017; Ravi Kumar & Ravi, 2007). In addition, discriminant analysis and logit regression provided a small number of selected variables because of strong correlations. These findings are consistent with those of Jabeur (2017), who reported that logit regression and discriminant analysis fail to resolve the problem of multicollinearity between ratios.

5.3 Analysis of Predictive Performance

Our study addresses a class imbalance problem. Approximately 90% of data observations are in the majority class (healthy firms), whereas only 10% of data observations are in the minority class (failed firms). Consequently, using accuracy rates leads to inconclusive results. According to Zhou et al. (2015), the area under the receiver operating characteristic curve (AUC) should be used as an accuracy measure. Following Berrar (2019), the AUC is defined as follows:

$$AUC = \int_0^1 TPR(t_i)dFPR(t_i)$$

where TPR(t_i) and FRP(t_i) denote the true positive rate and the false positive rate, respectively. Table 3 summarizes the AUC by period for each model estimated using a different feature selection technique: stepwise discriminant analysis, stepwise logistic regression, PLS-DA, or XGBoost (based on the most important variables). Tanagra[1] software by Rakotomalala (2005) was used for discriminant analysis, logit regression, support vector machines, partial least squares discriminant analysis, and Multi-Layer

[1] Tutorials explaining the methodology used by TANAGRA are available at the following address: https://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html.

Table 3 Area under the receiver operating characteristic curve (AUC) by type of feature selection method

Period	DA	LR	SVM	PLS-DA	NN	DNN	XGBoost
XGBoost feature selection models							
1Year	0.938	0.911	0.943	0.937	0.957	0.931	0.958
2Years	0.861	0925	0.888	0.864	0.891	0.918	0.931
3Years	0.799	0.826	0.795	0.795	0.825	0.798	0.866
	Stepwise logistic regression feature selection models						
1Year	0.922	0.881	0.921	0.921	0.935	0.857	0.935
2Years	0.871	0.869	0.883	0.874	0.910	0.931	0.929
3Years	0.809	0.814	0.756	0816	0.823	0.770	0.826
	Stepwise discriminant analysis feature selection models						
1Year	0.909	0.919	0.924	0.926	0.918	0.892	0.940
2Years	0.878	0.869	0.874	0.886	0.866	0.920	0.911
3Years	0.816	0.811	0.766	0.825	0.814	0.770	0.827
	Partial least squares discriminant analysis feature selection models						
Period	DA	LR	SVM	PLS-DA	NN	DNN	XGBoost
1Year	0.929	0.919	0.924	0.926	0.927	0.920	0.940
2Years	0.869	0.889	0.869	0.861	0.923	0.906	0.928
3Years	0.826	0.818	0.833	0.809	0.833	0.797	0.829

Rows labeled “1 year” refer to data 1 year prior to business failure, “2 years” data 2 years prior to failure, and “3 years” data 3 years prior to failure

Perceptron. Weka software by Lang et al. (2019) was used to run deep neural networks with the package WekaDeeplearning4j. XGBoost was fitted in R (R Core Team, 2019).

The predictive power was weaker for models used to predict corporate failure on more distant horizons. Similar findings have often been reported in previous studies (du Jardin, 2015, 2018). Overall, the results show that XGBoost based on the feature importance selection method (FS-XGBoost) performs better than stepwise discriminant analysis, stepwise logistic regression, and PLS-DA. In particular, FS-XGBoost outperformed all other models with different feature selection techniques. For the 1-year-ahead forecast, the AUC of FS-XGBoost ranged from 0.866 to 0.958. This was followed by PLS-DA and Multi-Layer Perceptron. Logistic regression yielded the worst results. In addition, XGBoost with traditional feature selection methods provided better AUC results than discriminant analysis, logit regression, support vector machines, partial least squares discriminant analysis, Multi-Layer Perceptron, and Restricted Boltzmann Machine. For example, the AUC of XGBoost based on PLS-DA feature selection ranged from 0.829 to 0.940. Hence, XGBoost based on PLS-DA feature selection outperformed deep neural networks, which had an AUC that ranged from 0.797 to 0.920. These results show the advantage over traditional models when XGBoost is combined with feature selection techniques.

Deep neural networks based on XGBoost feature selection provided the highest AUC for the different periods before failure. These results support those of Son et al.

(2019), who reported that XGBoost-based methods are superior to traditional techniques. Discriminant analysis and support vector machines combined with XGBoost feature selection performed significantly better than conventional techniques. Logit regression and Multi-Layer Perceptron combined with XGBoost feature selection did not provide significantly larger AUC values than stepwise discriminant analysis, stepwise logistic regression, or PLS-DA. The best combination of feature selection techniques and classification models one, two, and three years before failure was XGBoost based on the 10 most important variables (FS-XGBoost).

To provide more insightful findings, Table 4 presents the differences between the results calculated with XGBoost based on feature importance (FS-XGBoost) models and those estimated with all other models. The results are given by period and by other feature selection methods. The results show that FS-XGBoost systematically performs better than alternatives calculated with traditional feature selection methods. The differences range from 0.00 to 11.00 percentage points. In addition, the p-values of a test for differences between proportions indicate that more than three quarters of all differences are significant at a threshold lower than 10% (Table 4). This result emphasizes that model performance can be improved if the feature selection technique is chosen judiciously.

6 Conclusions

We propose a new forecasting financial distress approach that combines feature selection techniques and XGBoost (FS-XGBoost). Our results show that this model is more efficient than traditional feature selection techniques. The effectiveness appears to be supported by the solid ability of the approach to discriminate the observations. Using AUC as a measure of forecast accuracy, our tests reveal that FS-XGBoost was associated with a stronger discrimination power compared to other feature selection techniques, such as stepwise discriminant analysis, stepwise logistic regression or PLS-DA.

Two main conclusions may be drawn from this study. First, given its ability to identify the symptoms that lead to insolvency, FS-XGBoost provides a suitable alternative for corporate failure modeling, primarily for banks and financial institutions. Second, our estimates suggest that improvement in the AUC requires rigorous examination of different feature selection techniques. A reduction in the number of bankruptcy prediction variables usually leads to shorter processing times and lower levels of complexity.

Future studies should explore new feature selection methods that can be compared with traditional techniques. As pointed out by Stef and Jabeur (2018), one of the main goals of bankruptcy prediction is to identify the most accurate model that can correctly identify the largest number of firms for which a liquidation procedure was actually triggered. The achievement of such a goal should encourage researchers to address or propose new feature selection algorithms as an important step to identify the most significant features from a given dataset (Tsai, 2009). In addition, managers may view feature selection together with XGBoost algorithm as a methodology that can be used for the identification of the most relevant variables to evaluate the risk of business

Table 4 Percentage point differences between AUC with FS-XGBoost and AUC with other feature selection techniques

	Improvement			p-values of test for differences		
	1 year	2 years	3 years	1 year	2 years	3 years
Stepwise logistic regression feature selection models						
FS-XGBoost						
DA	3.6**	6.0***	5.7**	0.039	0.006	0.036
LR	7.7***	6.2***	5.2*	0.000	0.005	0.054
SVM	3.7**	4.8**	11.0***	0.035	0.025	0.000
PLS-DA	3.7**	5.7***	5.0*	0.035	0.009	0.063
NN	2.3	2.1	4.3	0.164	0.291	0.107
DNN	10.1***	0.0	9.6***	0.000	1.00	0.001
XGBoost	2.3	0.2	4	0.164	0.915	0.132
Stepwise discriminant analysis feature selection models						
FS-XGBoost						
DA	4.9***	5.3**	5.0*	0.007	0.014	0.063
LR	3.9**	6.2***	5.5**	0.029	0.005	0.042
SVM	3.4**	5.7***	10.0***	0.05	0.009	0.000
PLS-DA	3.2*	4.5**	4.1	0.063	0.034	0.123
NN	4.0**	6.5***	5.2*	0.024	0.003	0.054
DNN	6.6***	1.1	9.6***	0.001	0.569	0.001
XGBoost	1.8	2.0	3.9	0.266	0.313	0.141
Partial least squares discriminant analysis feature selection models						
FS-XGBoost						
DA	2.9*	6.2***	4.0	0.088	0.005	0.132
LR	3.9**	4.2**	4.8*	0.027	0.046	0.073
SVM	3.4**	6.2***	3.3	0.050	0.005	0.209
PLS-DA	3.2*	7.0	5.7**	0.063	0.002	0.036
NN	3.1*	0.8	3.3	0.07	0.676	0.209
DNN	3.8**	2.5	6.9**	0.031	0.214	0.012
XGBoost	1.8	0.3	3.7	0.266	0.873	0.162

Coefficients are estimated using XGBoost based on feature importance selection (FS-XGBoost). Columns labeled “1Year” refer to data 1 year prior to business failure, “2Years” data 2 years prior to failure, and “3Years” data 3 years prior to failure

*At the 10% level

**At the 5% level

***Denotes a significant coefficient at the 1% level failure. In the light of the high predictive power of the model developed in this paper, business managers should be encouraged to identify the most important features and complete a parsimonious model that can anticipate firm's financial problems.

This study might yield some interesting research potentials. First, expanding the sample to include a broader scope could improve the model's strength and assure its generalization to other countries, for example considering supranational data in the context of E.U. In addition, taking a balanced sample data would help corroborating the obtained results. Second, FS-XGBoost relies on a binary dependent variable. However, Stef and Zenou (2021) argued that firm's exit can follow multiple paths such as court-driven exit or out-of-court exit. By examining the travel mode choices, Wang and Ross (2018) found that the XGBoost model has higher prediction accuracy than the multinomial logit model in the case of extremely unbalanced dataset. Therefore, the treatment of different exit paths through FS-XGBoost can open new research opportunities. Third, Gilbert et al. (1990) suggested that bankruptcy modelling should not be based solely on financial ratios. Consequently, the accuracy of FS-XGBoost should be investigated by considering additional nonfinancial variables such as firm's ownership, managers' experience, local competition or consumer segments.

Funding The authors received financial support from the Spanish Ministry of Science, Innovation and Universities. FEDER project PGC2018-093645-B-I00 is gratefully acknowledged.

References

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *23*(4), 589–609.

Alzubi, J. A., Bharathikannan, B., & Tanwar, S. (2019). Boosted neural network ensemble classification for lung cancer disease diagnosis. *Applied Soft Computing Journal, 80*, 579–591. https://doi.org/10.1016/j.asoc.2019.04.031

Ander, J., Arévalo, J., Paredes, R., & Nin, J. (2018). End-to-end neural network architecture for fraud scoring in card payments. *Pattern Recognition Letters, 105*, 175–181. https://doi.org/10.1016/j.patrec.2017.08.024

Bao, W., Lianju, N., & Yue, K. (2019). Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. *Expert Systems with Applications, 128*, 301–315. https://doi.org/10.1016/j.eswa.2019.02.033

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

Bardos, M. (1998). Detecting the risk of company failure at the Banque de France. *Journal of Banking & Finance, 22*(10–11), 1405–1419. https://doi.org/10.1016/S0378-4266(98)00062-4

Bastien, P., Vinzi, V. E., & Tenenhaus, M. (2005). PLS generalised linear regression. *Computational Statistics and Data Analysis, 48*(1), 17–46. https://doi.org/10.1016/j.csda.2004.02.005

Becker, J. M., & Ismail, I. R. (2016). Accounting for sampling weights in PLS path modeling: Simulations and empirical examples. *European Management Journal, 34*(6), 606–617. https://doi.org/10.1016/j.emj.2016.06.009

Bellini, S., Cardinali, M. G., & Grandi, B. (2017). A structural equation model of impulse buying behaviour in grocery retailing. *Journal of Retailing and Consumer Services, 36*, 164–171.

Berrar, D. (2019). Performance measures for binary classification. In S. Ranganathan, M. Gribskov, K. Nakai, & C. Schönbach (Eds.), *Encyclopedia of Bioinformatics and Computational Biology* (pp. 546–560). Oxford: Academic Press. https://doi.org/10.1016/B978-0-12-809633-8.20351-8.

Blazy, R., & Stef, N. (2020). Bankruptcy procedures in the post-transition economies. *European Journal of Law and Economics, 50*(1), 7–64. https://doi.org/10.1007/s10657-019-09634-5

Bolón-Canedo, V., & Alonso-Betanzos, A. (2019). Ensembles for feature selection: A review and future trends. *Information Fusion, 52*, 1–12. https://doi.org/10.1016/j.inffus.2018.11.008

Boloukian, B., & Safi-esfahani, F. (2019). *Journal neural networks*. Elsevier. https://doi.org/10.1016/j.neunet.2019.07.012

Brezigar-Masten, A., & Masten, I. (2012). CART-based selection of bankruptcy predictors for the logit model. *Expert Systems with Applications, 39*(11), 10153–10159. https://doi.org/10.1016/j.eswa.2012.02.125

Buuren, S. V., & Groothuis-Oudshoorn, K. (2010). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*, 1–68.

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

Chang, Y., Chang, K., & Wu, G. (2018). Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Applied Soft Computing Journal, 73*, 914–920. https://doi.org/10.1016/j.asoc.2018.09.029

Chatzis, S. P., Siakoulis, V., Petropoulos, A., Stavroulakis, E., & Vlachogiannakis, N. (2018). Forecasting stock market crisis events using deep and statistical machine learning techniques. *Expert Systems with Applications, 112*, 353–371. https://doi.org/10.1016/j.eswa.2018.06.032

Chen, H. J., Huang, S. Y., & Lin, C. S. (2009). Alternative diagnosis of corporate bankruptcy: A neuro fuzzy approach. *Expert Systems with Applications, 36*(4), 7710–7720. https://doi.org/10.1016/j.eswa.2008.09.023

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In*Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Choi, H., Son, H., & Kim, C. (2018). Predicting financial distress of contractors in the construction industry using ensemble learning. *Expert Systems with Applications, 110*, 1–10. https://doi.org/10.1016/j.eswa.2018.05.026

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

Cordón, I., Luengo, J., García, S., Herrera, F., & Charte, F. (2019). Smartdata : Data preprocessing to achieve smart data in R. *Neurocomputing, 360*, 1–13. https://doi.org/10.1016/j.neucom.2019.06.006

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

du Jardin, P., & Séverin, E. (2012). Forecasting financial failure using a Kohonen map: A comparative study to improve model stability over time. *European Journal of Operational Research, 221*(2), 378–396. https://doi.org/10.1016/j.ejor.2012.04.006

Daoud, M., Mayo, M., Box, P. O., & Zealand, N. (2019). A survey of neural network-based cancer prediction models from microarray data. *Artificial Intelligence in Medicine, 97*, 204–214. https://doi.org/10.1016/j.artmed.2019.01.006

Dixon, M., Klabjan, D., & Bang, J. H. (2015). Implementing deep neural networks for financial market prediction on the Intel Xeon Phi. In *Proceedings of the 8th workshop on high performance computational finance—WHPCF '15* (pp. 1–6). https://doi.org/10.1145/2830556.2830562

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). Bankruptcy prediction using terminal failure processes. *European Journal of Operational Research, 242*(1), 286–303. https://doi.org/10.1016/j.ejor.2014.09.059

Everett, J., & Watson, J. (1998). Small business failure and external risk factors. *Small Business Economics, 11*(4), 371–390. https://doi.org/10.1023/A:1008065527282

Fernández-Gámez, M. Á., Cisneros-Ruiz, A. J., & Callejón-Gil, Á. (2016). Applying a probabilistic neural network to hotel bankruptcy prediction. *Tourism & Management Studies, 12*(1), 40–52. https://doi.org/10.18089/tms.2016.12104

Friedman, J. (2001). Greedy function approximation : A gradient boosting machine author (s): Jerome H . Friedman Source : The Annals of Statistics , Vol . 29 , No . 5 (Oct ., 2001), pp . 1189–1232 Published by : Institute of Mathematical Statistics Stable. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/009053606000000795

García, S., Luengo, J., & Herrera, F. (2016). Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Knowledge-Based Systems, 98*, 1–29. https://doi.org/10.1016/j.knosys.2015.12.006

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big data preprocessing: Methods and prospects. *Big Data Analytics, 1*(1), 9. https://doi.org/10.1186/s41044-016-0014-0

Geng, R., Bose, I., & Chen, X. (2015). Prediction of financial distress: An empirical study of listed Chinese companies using data mining. *European Journal of Operational Research, 241*(1), 236–247. https://doi.org/10.1016/j.ejor.2014.08.016

Gilbert, L. R., Menon, K., & Schwartz, K. B. (1990). Predicting bankruptcy for firms in financial distress. *Journal of Business Finance & Accounting, 17*(1), 161–171. https://doi.org/10.1111/j.1468-5957.1990.tb00555.x

Hernandez Tinoco, M., & Wilson, N. (2013). Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables. *International Review of Financial Analysis, 30*, 394–419. https://doi.org/10.1016/j.irfa.2013.02.013

Hinton, G. E. (2006). Communicated by Yann Le Cun A fast learning algorithm for deep belief nets 500 units 500 units. *Neural Computation, 1554*, 1527–1554.

Hu, Z., Tang, J., Wang, Z., Zhang, K., Zhang, L., & Sun, Q. (2018). Deep learning for image-based cancer detection and diagnosis—A survey. *Pattern Recognition, 83*, 134–149. https://doi.org/10.1016/j.patcog.2018.05.014

Jabeur, S. B. (2017). Bankruptcy prediction using Partial Least Squares Logistic Regression. *Journal of Retailing and Consumer Services, 36*, 197–202. https://doi.org/10.1016/j.jretconser.2017.02.005

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*, 120658. https://doi.org/10.1016/j.techfore.2021.120658

Jabeur, S. B., Sadaaoui, A., Sghaier, A., & Aloui, R. (2020). Machine learning models and cost-sensitive decision trees for bond rating prediction. *Journal of the Operational Research Society, 71*(8), 1161–1179. https://doi.org/10.1080/01605682.2019.1581405

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

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

Jardin, P. (2018). Failure pattern-based ensembles applied to bankruptcy forecasting. *Decision Support Systems, 107*, 64–77. https://doi.org/10.1016/j.dss.2018.01.003

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.

Kim, E., Lee, J., Shin, H., Yang, H., Cho, S., Nam, S., et al. (2019). Champion-challenger analysis for credit card fraud detection : Hybrid ensemble and deep learning. *Expert Systems with Applications, 128*, 214–224. https://doi.org/10.1016/j.eswa.2019.03.042

Kim, H. J., Jo, N. O., & Shin, K. S. (2016). Optimization of cluster-based evolutionary undersampling for the artificial neural networks in corporate bankruptcy prediction. *Expert Systems with Applications, 59*, 226–234. https://doi.org/10.1016/j.eswa.2016.04.027

Kim, M. J., & Kang, D. K. (2010). Ensemble with neural networks for bankruptcy prediction. *Expert Systems with Applications, 37*(4), 3373–3379. https://doi.org/10.1016/j.eswa.2009.10.012

Kraus, M., & Feuerriegel, S. (2017). Decision support from financial disclosures with deep neural networks and transfer learning. *Decision Support Systems, 104*, 38–48. https://doi.org/10.1016/j.dss.2017.10.001

Krawczyk, B., & Herrera, F. (2017). A survey on data preprocessing for data stream mining: Current status and future directions. *Neurocomputing, 239*, 39–57. https://doi.org/10.1016/j.neucom.2017.01.078

Kumar, A., Ramachandran, M., Gandomi, A. H., & Patan, R. (2019). A deep neural network based classifier for brain tumor diagnosis. *Applied Soft Computing Journal, 82*, 105528. https://doi.org/10.1016/j.asoc.2019.105528

Laitinen, E. K., Lukason, O., & Suvas, A. (2014). Are firm failure processes different? Evidence from seven countries. *Investment Management and Financial Innovations, 11*(4), 212–222.

Lang, S., Bravo-marquez, F., Beckham, C., Hall, M., & Frank, E. (2019). WekaDeeplearning4j : A deep learning package for Weka based on. *Knowledge-Based Systems, 178*, 48–50. https://doi.org/10.1016/j.knosys.2019.04.013

Lee, K., Booth, D., & Alam, P. (2005). A comparison of supervised and unsupervised neural networks in predicting bankruptcy of Korean firms. *Expert Systems with Applications, 29*(1), 1–16. https://doi.org/10.1016/j.eswa.2005.01.004

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

Leong, L., Hew, T., Tan, G. W., & Ooi, K. (2013). Predicting the determinants of the NFC-enabled mobile credit card acceptance : A neural networks approach. *Expert Systems with Applications, 40*(14), 5604–5620. https://doi.org/10.1016/j.eswa.2013.04.018

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

Liang, D., Tsai, C. F., & Wu, H. T. (2014). The effect of feature selection on financial distress prediction. *Knowledge-Based Systems, 73*(1), 289–297. https://doi.org/10.1016/j.knosys.2014.10.010

Liang, D., Tsai, C. F., & Wu, H. T. (2015). The effect of feature selection on financial distress prediction. *Knowledge-Based Systems, 73*(1), 289–297. https://doi.org/10.1016/j.knosys.2014.10.010

Lin, F., Liang, D., Yeh, C. C., & Huang, J. C. (2014). Novel feature selection methods to financial distress prediction. *Expert Systems with Applications, 41*(5), 2472–2483. https://doi.org/10.1016/j.eswa.2013.09.047

Mai, F., Tian, S., Lee, C., & Ma, L. (2019). Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research, 274*(2), 743–758. https://doi.org/10.1016/j.ejor.2018.10.024

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

Olson, D. L., Delen, D., & Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems, 52*(2), 464–473. https://doi.org/10.1016/j.dss.2011.10.007

Platt, H. D., & Platt, M. B. (1994). Business cycle effects on state corporate failure rates. *Journal of Economics and Business*, *46*(2), 113–127.

Qawaqneh, Z., Abu, A., & Barkana, B. D. (2017). Age and gender classification from speech and face images by jointly fine-tuned deep neural networks. *Expert Systems with Applications, 85*, 76–86. https://doi.org/10.1016/j.eswa.2017.05.037

R Core Team. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria (2019). http://www.R-project.org/.

Rakotomalala, R. (2005) TANAGRA: A free software for research and academic purposes. In *Proceedings of EGC'2005, RNTI-E-3* (Vol. 2, pp. 697–702).

Ravi Kumar, P., & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques—A review. *European Journal of Operational Research, 180*(1), 1–28. https://doi.org/10.1016/j.ejor.2006.08.043

Ravisankar, P., & Ravi, V. (2010). Financial distress prediction in banks using Group Method of Data Handling neural network, counter propagation neural network and fuzzy ARTMAP. *Knowledge-Based Systems, 23*(8), 823–831. https://doi.org/10.1016/j.knosys.2010.05.007

Serrano-cinca, C., & Gutiérrez-nieto, B. (2013). Partial least square discriminant analysis for bankruptcy prediction. *Decision Support Systems, 54*(3), 1245–1255. https://doi.org/10.1016/j.dss.2012.11.015

Sghaier, A., Ben Jabeur, S., & Bannour, B. (2018). Using partial least square discriminant analysis to distinguish between Islamic and conventional banks in the MENA region. *Review of Financial Economics, 36*(2), 133. https://doi.org/10.1002/rfe.1018

Shi, X., Wong, Y. D., Li, M.Z.-F., Palanisamy, C., & Chai, C. (2019). A feature learning approach based on XGBoost for driving assessment and risk prediction. *Accident Analysis & Prevention, 129*, 170–179. https://doi.org/10.1016/j.aap.2019.05.005

Shin, K. S., Lee, T. S., & Kim, H. J. (2005). An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications, 28*(1), 127–135. https://doi.org/10.1016/j.eswa.2004.08.009

Son, H., Hyun, C., Phan, D., & Hwang, H. J. (2019). Data analytic approach for bankruptcy prediction. *Expert Systems with Applications, 138*, 112816. https://doi.org/10.1016/j.eswa.2019.07.033

Stef, N. (2018). Bankruptcy and the difficulty of firing. *International Review of Law and Economics, 54*, 85–94. https://doi.org/10.1016/j.irle.2017.11.002

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

Stef, N., & Jabeur, S. B. (2018). The bankruptcy prediction power of new entrants. *International Journal of the Economics of Business, 1516*, 1–20. https://doi.org/10.1080/13571516.2018.1455389

Stef, N., & Zenou, E. (2021). Management-to-staff ratio and a firm's exit. *Journal of Business Research, 125*, 252–260. https://doi.org/10.1016/j.jbusres.2020.12.027

Tsai, C. F. (2009). Feature selection in bankruptcy prediction. *Knowledge-Based Systems, 22*(2), 120–127. https://doi.org/10.1016/j.knosys.2008.08.002

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

Tsakonas, A., Dounias, G., Doumpos, M., & Zopounidis, C. (2006). Bankruptcy prediction with neural logic networks by means of grammar-guided genetic programming. *Expert Systems with Applications, 30*(3), 449–461. https://doi.org/10.1016/j.eswa.2005.10.009

Wang, F., & Ross, C. L. (2018). Machine learning travel mode choices: Comparing the performance of an extreme gradient boosting model with a multinomial logit model. *Transportation Research Record: Journal of the Transportation Research Board, 2672*(47), 35–45. https://doi.org/10.1177/0361198118773556

Wang, G., Ma, J., & Yang, S. (2014). An improved boosting based on feature selection for corporate bankruptcy prediction. *Expert Systems with Applications, 41*(5), 2353–2361. https://doi.org/10.1016/j.eswa.2013.09.033

Wold, H. (1985). Partial least squares. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 6, pp. 581–591). Wiley.

Wruck, K. H. (1990). Financial distress, reorganization, and organizational efficiency. *Journal of Financial Economics, 27*(2), 419–444. https://doi.org/10.1016/0304-405X(90)90063-6

Xia, Y., Liu, C., Li, 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

Yang, Z., You, W., & Ji, G. (2011). Using partial least squares and support vector machines for bankruptcy prediction. *Expert Systems with Applications, 38*(7), 8336–8342. https://doi.org/10.1016/j.eswa.2011.01.021

Yu, J., Shi, S., Zhang, F., Chen, G., & Cao, M. (2019). PredGly: Predicting lysine glycation sites for Homo sapiens based on XGboost feature optimization. *Bioinformatics, 35*(16), 2749–2756. https://doi.org/10.1093/bioinformatics/bty1043

Zhang, H., Qiu, D., Wu, R., Deng, Y., Ji, D., & Li, T. (2019). Novel framework for image attribute annotation with gene selection XGBoost algorithm and relative attribute model. *Applied Soft Computing Journal, 80*, 57–79. https://doi.org/10.1016/j.asoc.2019.03.017

Zhang, R., Nie, F., Li, X., & Wei, X. (2019). Feature selection with multi-view data: A survey. *Information Fusion, 50*, 158–167. https://doi.org/10.1016/j.inffus.2018.11.019

Zhao, D., Huang, C., Wei, Y., Yu, F., Wang, M., & Chen, H. (2017). An effective computational model for bankruptcy prediction using kernel extreme learning machine approach. *Computational Economics, 49*(2), 325–341. https://doi.org/10.1007/s10614-016-9562-7

Zhou, L., & Lai, K. K. (2017). AdaBoost models for corporate bankruptcy prediction with missing data. *Computational Economics, 50*(1), 69–94. https://doi.org/10.1007/s10614-016-9581-4

Zhou, L., Lu, D., & Fujita, H. (2015). The performance of corporate financial distress prediction models with features selection guided by domain knowledge and data mining approaches. *Knowledge-Based Systems, 85*, 52–61. https://doi.org/10.1016/j.knosys.2015.04.017

Zhou, L., Si, Y., & 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