

ANTICIPATING CORPORATE'S DISTRESSES

ANTECIPANDO DIFICULDADES FINANCEIRAS NAS ORGANIZAÇÕES

Recebido em: 22 jun. 2020


Aprovado em: 18 out. 2020

Versão do autor aceita publicada online: 18 out. 2020

Publicado online: 11 jun. 2021

Como citar esse artigo - American Psychological Association (APA):

Barboza, F. L. de M., Duarte, D. L., & Cunha, M. A. (2022, abr./jun.). Anticipating corporate's distresses. *Exacta*. 20(2), 470-496. <https://doi.org/10.5585/exactaep.2021.17494>.

Submeta seu artigo para este periódico 



Dados Crossmark



ANTICIPATING CORPORATE'S DISTRESSES

ANTECIPANDO DIFICULDADES FINANCEIRAS NAS ORGANIZAÇÕES

 Flávio Luiz de Moraes Barboza¹

 Denize Lemos Duarte²

 Michele Aparecida Cunha³

Abstract:: The objective of this study is to present a model for predicting Financial Distress (FD) from the perspective of machine learning techniques (TAMs). We applied and compared the XGBoost, Random Forest and Logistic Regression models using financial indicators to seek better forecasts of FDs one year before the event in Latin American companies in the period from 2000 to 2017. Our results showed that TAMs outperform the logit model, reaching an overall accuracy of 96% (XGBoost). In addition, five indicators were relevant to its success. The study expands knowledge and discussions by focusing on the predictive power of comparing models, highlighting the benefits of using algorithms applied to financial research. It helps in risk management, loss prevention, allowing greater balance and health for the financial system, which contributes to the economic, social and sustainable development of a society.

Keywords: Credit Risk. Financially distressed company. Artificial intelligence. Extreme gradient boosting. Variable importance.

Resumo: O objetivo deste estudo é apresentar um modelo de previsão de Dificuldades Financeiras (DF) a partir da perspectiva das técnicas de aprendizado de máquina (TAMs). Aplicamos e comparamos os modelos XGBoost, Random Forest e Regressão Logística usando indicadores financeiros para buscar melhores previsões das DFs um ano antes do evento em empresas latino-americanas no período de 2000 a 2017. Nossos resultados mostraram que as TAMs superam o modelo de logit, atingindo uma precisão geral de 96 % (XGboost). Além disso, cinco indicadores foram relevantes para o seu sucesso. O estudo amplia o conhecimento e as discussões ao focar o poder preditivo na comparação entre os modelos, destacando os benefícios do uso de algoritmos aplicados à pesquisa financeira. Auxilia na gestão de riscos, na prevenção de perdas, permitindo maior equilíbrio e saúde para o sistema financeiro, que contribui para o desenvolvimento econômico, social e sustentável de uma sociedade.

Palavras-chave: Risco de Crédito. Dificuldade financeira corporativa. Inteligência artificial. Extreme gradient boosting. Importância de variáveis.

¹ Doutor em Finanças Estratégicas
Universidade Federal de Uberlândia – UFU
Uberlândia, Minas Gerais - Brasil
flmbarboza@ufu.br

² Mestre em Administração
Universidade Federal de Uberlândia – UFU
Uberlândia, Minas Gerais – Brasil
denize_ld@hotmail.com

³ Mestre em Administração
Universidade Federal de Uberlândia – UFU
Uberlândia, Minas Gerais – Brasil
micheleapcunha@hotmail.com

Recebido em: 22 jun. 2020

Aprovado em: 18 out. 2020



1 Introduction


The prediction of financial distress (FD) is widely studied in corporate finance because of its impact on the very survival and development of the company, as well as on the decision of external investors and creditors, as it is an effective way of managing risks (Tang, Li, Tan & Shi, 2020). In turn, when companies are unable to pay their obligations due to financial distress, the chances of banks not receiving their money back increase and this would cause problems for the entire financial system (Sun, Fujita, Chen & Li, 2017).

Sun et al. (2017) state that in situations like these it is necessary to develop efficient FD forecasting models, which can help both companies to improve risk management, and banks to make more accurate and assertively credit decisions. The development of models that are capable of predicting default and or FD are crucial for both (creditors and debtors) to be able to take both preventive and corrective actions (Wang, Wang & Lai, 2005; Lai, Yu, Wang & Zhou, 2006).

In a large study by Jones, Johnstone & Wilson (2015) compared the performance of traditional models (Logistic Regression and Discriminant Analysis) to machine learning models, such as Neural Networks, Support Vector Machine (SVM) and more recent techniques such as generalized boosting, AdaBoost and Random Forest. In this article, the authors demonstrated that the latter outperformed all other methods.

In addition to its efficiency in data analysis, machine learning has aspects that should be researched with more focus to encourage its adoption in companies. Wang & Ma (2011) pointed out that the models must combine precision and usability. Chen, Yang, Wang, Liu, Xu, Wang & Liu (2011) suggested improving the interpretability of ensembles as an important research direction, due to the lack of satisfactory conclusions. Bae (2012) recommended exploring and building models with different data sets, as this is a delicate problem, even the banking market may offer restrictions on the disclosure of its customers' information.

Scientific studies are contributing in different ways. In particular, Guo, Zhou, Luo, Liu & Xiong (2016) asserted that the models are based only on numerical and financial variables. Then, they recommended the use of non-financial variables as factors related to corporate governance (for example, management capacity, reputation, type of ownership, plans, etc.), macroeconomic conditions of corporate and consumer performance, and even social media data. Tang et al. (2020) used non-financial characteristics to predict financial distress for Chinese companies while also incorporating management and performance factors in different periods. As a consequence, the authors realized that management and non-financial factors can complement the data in the forecasting statements and that the factors collected four years before the reference year allowed a more accurate assessment of the forecast.



Zięba, Tomczak & Tomczak (2016) used statistical and artificial intelligence-based methods such as SVM, neural networks and Extreme Gradient Boosting (XGBoost), for bankruptcy forecasting with financial data from Poland during 2000–2012. XGBoost's results were better than the other reference methods with significant evolution in the quality of the forecast.

Chang & Hsu (2018) employed XGBoost to build a credit risk assessment model for financial institutions with Taiwanese loan data about 8 years. In this work, the model's precision was superior to SVM, Group Method of Data Handling (GMDH) and Logit. Climent, Momparler & Carmona (2019) applied XGBoost to identify the influence of indicators and variables in the forecast of bank failures with Eurozone data for the period 2006–2016 for testing the predictive power of the model, which presented a satisfactory result for the banking sector. Therefore, they examined the XGBoost technique, since it is a recent machine learning method used for supervised learning problems (Chen & Guestrin, 2016), considered an improvement and based on the Friedman model 2001. Yet, this method represents an advance in the use of new techniques to predict bankruptcies, as made by Zięba et al. (2016) and FD (Huang & Yen, 2019), which are potentially indicative scenarios that your performance is successful in the context of our research, which is to predict financial distress for Latin American companies.

The statistical method based on artificial intelligence Extreme Gradient Boosting (Xgboost) is a recent machine learning method used for supervised learning problems that shows superiority of precision over others. (CHEN; GUESTRIN, 2016). Kim & Kang (2010), Finlay (2011), Brown & Mues (2012), Tsai, Hsu & Yen (2014) and Kim, Kang & Kim (2015) recommended the development of models taking into account the methods of boosting and bagging, which is based on a constructive analysis strategy. Thus, from the gaps pointed out by these studies, these recommendations were observed to continue research on the topic.

The choice of this model rests on demonstrated efficiency, speed, precision and practicality of its algorithm (Chang & Hsu, 2018). In addition, its ability to perform various calculations on bases with an extensive volume of data, even on an ordinary computer, makes the choice of this model even more promising. The algorithm is built in such a way that it also has additional resources to perform cross-validation and be able to display the most impacting variables. In addition to data analysis, made some other interesting efforts on the XGBoost algorithm, such as memory optimization, cache optimization and improvement in terms of the model itself, with external memory programming, distributed abstractions, which helps to understand which algorithm it is suitable for which path.

Another gap observed for the development of this study is the scarcity of credit risk surveys in Latin America. Our survey of articles on the topic revealed the main articles with the highest number of citations and, therefore, those of greatest scientific relevance, and we identified that most of the relevant studies on the topic come from China and Taiwan, these studies are the basis of reference for others carried out in Europe, North America and Oceania. However, the regions of Latin America and



Africa have a shortage of relevant works on the topic. Therefore, studies covering information from these regions should produce interesting results.

When it comes to models for Latin America, it is necessary to expand the variables and measures to be evaluated, as it deals with a complex market, with economic, political and cultural inequalities, which motivate a greater analysis of the factors that may contribute to the development of that region (Martinez-Villa & Machin-Mastromatteo, 2016).

In this context, considering the relevance of innovation in the management of credit risk in Latin America, the objective of this article is to develop and compare the predictive power of three techniques to predict financial distress and bankruptcies. For this, our sample consisted of companies listed on the stock exchange, as all information is public and available. Given the advantages and disadvantages presented by the literature of some techniques,


this study seeks to test the XGBosst classification model to determine the likelihood of Latin American companies reaching the FD stage in the next year. In addition, we also verify which variables should be considered to anticipate the event. Wang & Ma (2011) applied RS-Boosting and obtained better results, mainly in the reduction of type II error, but from their work showed that the interpretability of the results is another important research direction that is still little explored.

The main objective is not only to oppose different methods of machine learning, but also to highlight a recently developed model and, thus, to motivate research uniting areas of finance and computer science. We compared the XGBoost model with a conventional statistical method (logistic regression) and a well-known machine learning algorithm (Random Forest). Complementary to its performance, XGBoost has another important aspect which is the identification and display of the most important variables for the model.

Although there are previous studies in the literature, it is essential to treat other poorly studied regions, such as Latin America, which has different market attributes, so that each country or region has a model for predicting bankruptcies and financial distress, taking into account the structure of the system domestic financial and market peculiarities (Korol, 2013).

We developed our empirical analysis using data from Latin America with more than a thousand companies in the period between 2000 and 2017. The results of our research suggest that the performance and precision of the XGBoost model, when compared to the most used models, such as Regression Logistics and random forest, is superior. XGBoost achieved an accuracy rate of 96.05%, versus 95.10% of the random forest and 65.12% of the logistic regression.

These findings contribute to the literature on predicting credit risk in some ways. First, this becomes an important fact, because the codes of the algorithm are available to the public so that banks and future investors, analysts and managers can avoid the pre-concept "black box" that complex models have in practical applications. Second, this research aims to encourage the alliance between finance and



data science in the search for a structured and accurate decision-making process, with a focus on risk mitigation and better quality in the analyzes that precede the two strategic decisions in terms of resources and competitiveness) and operational (short and medium term, in direct contact with the client).

However, this research has interesting implications. Among them, they highlight the possibility of readjusting the risk classifications, bringing more accurate information about the effective quality of the insured's risk and, thus, adjusting the spreads accordingly. Another relevant implication is the improvement of risk management in companies in different sectors, as they will be able to observe and even argue with creditors their exposure to credit risk using the necessary skill.

In addition to this introduction, this article is organized as follows: Section 2 provides a review of the literature on credit risk and forecasting models. Section 3 presents the research methodology. Section 4 examines the empirical findings regarding the XGBoost model. Finally, section 5 presents the conclusions and suggestions for future research.

2 Related literature

According to Altman (1968), in order to detect the solvency of a firm, it is essential to analyze financial ratios to predict and measure the payment potential of that company. Beaver (1966) defines FD as a firm's inability to pay its financial obligations and points out that the usefulness of financial ratios is measured by the underlying predictive capacity of accounting data.

The Basel II Agreement requires companies to disclose risk management practices, which requires more reliable and accurate models to classify and quantify these probabilities (BIS, 2006). For this reason, the adoption of machine learning algorithms, a subtopic of data sciences that refers to the study of pattern recognition theory and computational learning using artificial intelligence.

Studies have compared the performance of different methods. Alfaro, García, Gámez & Elizondo (2008) confirmed that AdaBoost outperforms Neural Networks and that their error test was 8.898% against 12.712% of neural networks. Heo & Yang (2014) compared several success rates of the machine learning algorithm and tested it against Altman's famous Z-Score Altman (1968): AdaBoost (78.5%), ANN: (77.1%) SVM (73.3%), DT (73.1%) and Altman's Z-Score (51.3%).

Sung, Chang & Lee (1999) developed bankruptcy prediction models (discriminant analysis, neural networks, decision trees and genetic algorithms) with data from Korea adapting the models to normal and crisis economic conditions, providing a possible interpretation of the causes bankruptcy without losing the perceived predictive ability of each model.



Korol (2013) investigated and compared statistical methods (discriminant analysis and decision trees) and computational methods (neural networks) for Latin and Central America and European companies. The author used data from 185 companies listed on the Warsaw bold list and 60 companies listed in Mexico, Argentina, Peru, Brazil and Chile. The results were highly effective in forecasting, demonstrating that it is possible to efficiently predict bankruptcy in companies in dynamic and unstable markets such as those in Latin America. In addition, the results of the decision tree models were more promising than the most popular methods of discriminant analysis and artificial neural networks.

Barboza, Kimura & Altman (2017) tested machine learning models (vector machines, bagging, reinforcement and random forest) to predict bankruptcy and bought using models of discriminant analysis, logistic regression and neural networks. The authors used data from 1985 to 2013 in North American companies, integrating data from Salomon Center and Compustat. The authors concluded that bagging, boosting, and random forests outperformed techniques such as Logit, Discriminant Analysis, and Support Vector Machine.


Huang & Yen (2019) made a comparison between six machine learning models in the context of forecasting financial distress, using data from financial statements of Taiwanese companies from 2010 to 2016. The authors demonstrated that among the supervised algorithms, the XGBoost provided the most accurate forecast. Carmona, Climent & Momparler (2019) used XGBoost to predict bankruptcies in the US banking sector and also presented an interpretation of the model's outputs, providing information on the variables used. The result of the model was that predictive power is greater than most conventional methods. In Table 1, it can be seen that researchers around the world have struggled with different algorithms. Such studies highlight the capacity of the models, but also point out their disadvantages, such as their obscure nature, greater computational burden, the propensity to overfitting and empirical nature of the construction.

Table 1

Machine learning methods used in credit risk

Technique	References of Application
Boosting	Sun et al. (2011); Wang & Ma (2011); Wang, Ma & Yang (2014); Kim & Upneja (2014); Heo & Yang (2014); Kim et al. (2015); Sun et al. (2017); Zięba et al. (2016); Chang & Hsu (2018); Carmona et al. (2019); Climent et al. (2019); Huang & Yen (2019); Tang et al. (2020).
Neural Networks	Tsai & Wu (2008); Kim & Kang (2010); Chuang & Huang (2011); Jeong, Min & Kim (2012); Blanco, Pino-Mejías, Lara & Rayo (2013); Korol (2013); Iturriaga & Sanz (2015); Zhao et al. (2015); Yu, Yang & Tang (2016).
SVM	Li & Sun (2012); Wang & Ma (2011); Hens & Tiwari (2012); Hsieh, Hsiao & Yeh (2012); Harris (2015); Danenas & Garsva (2015); Sun et al. (2017).
Decision Trees	Li et al. (2010b); Cho, Hong & Ha (2010); Zhang, Zhou, Leung & Zheng (2010); Gepp, Kumar & Bhattacharya (2010); Wang & Ma (2011); Kim & Upneja (2014).

Source: Organized by the authors.



Although almost all methods can be used to assess credit risk, recently - due to the increasing complexity and size of databases - researchers have experimented with different classifiers and techniques, which integrate two or more classification methods. These approaches have shown greater precision in predictability than individual methods and have attracted a great deal of attention in credit risk assessment. Some of these examples are the Neural Discriminating Technique (Lee, Chiu, Lu & Chen, 2002), neuro-fuzzy (Piramuthu, 1999; Malhotra & Malhotra, 2002) and fuzzy-SVM (Wang et al., 2005).

In a database with financial information from customers, there is a scenario in which the number of observations associated with one class is much greater than those belonging to the other class. For this reason, there is an issue of information disproportion, which Brown & Mues (2012) examined in their study investigating various types of algorithms, the results of which demonstrated that Random Forest - an algorithm based on decision trees - and Gradient Boosting performed moderately well on unbalanced data sets.

In finance, the application of this technique is relatively new. He, Zhang & Zhang (2018) compared XGBoost's performance with other credit scoring models and obtained the best ranking results in four of the six databases used, indicating that it has a consistent potential for better performance compared to other techniques. Carmona et al. (2019) tested the model to predict bankruptcies in the US banking sector and concluded that XGBoost has greater predictive power than the Logistic Regression and Random Forest methods. Xia, Liu, Da & Xie (2018) highlights the model's superiority as a meta-classifier, that is, a high-level learning algorithm. Xia, Liu, Li & Liu (2017) highlighted comparisons with different base models and showed the superiority of the model based on XGBoost in terms of predictive performance.

In contrast to algorithms that use the bagging method, which builds models in parallel, the boosting approach is to build models sequentially. The XGBoost uses a certain number of classifiers to approach the final model and minimize the given cost function. The gradient descent method calculates the partial derivative and tries to optimize the cost function by looking for the (local) minimum by adjusting different coefficient values to iteratively minimize the error. This cost function measures how well the model fits current data and the boosting process continues until the reduction of the cost function becomes limited (Chen & Guestrin, 2016).

3 Sample and methodology

3.1 The Database

The research sample is composed of all Latin American companies with available data in the Economatica database, covering the period 2000–2017. These data include canceled stocks (e.g., delisted) to mitigate survival bias, otherwise the sample would be much smaller if only active companies



were considered at the end of 2017 (Iquipaza, Lamounier & Amaral, 2008). Companies in the financial sector were excluded from the sample due to the specific accounting and sector characteristics, in accordance with Campos & Nakamura (2015). The number of companies in the final sample: active companies: 702; (+) companies canceled: 353; (=) final sample: 1,055 companies.

The variable of interest represents a label according to Pindado, Rodrigues & De La Torre (2008), which measures the company's ability to manage its debt. Thus, this study uses a financial criterion to determine this variable, considering a definition of financial distress (FD) based on the company's failure to meet its obligations is consistent with an ex-ante approach (Sanz & Ayca, 2006).

We classify a company as FD when both of the following conditions are satisfied: 1) earnings before interest and taxes, depreciation and amortization (EBITDA) are lower than its financial expenses for two consecutive years, leading the company to a situation in which it cannot generate sufficient funds from its operating activities to meet its financial obligations; (2) a fall in its market value for two consecutive periods (Manzaneque, Priego & Merino, 2016). Therefore, this research considers a company with FD in the year that immediately follows the occurrence of these two events. This condition is represented by a binary dependent variable that has a value of 1 (one) for companies with FD and 0 otherwise.


Regarding the independent variables, we identified potential variables in the literature to predict DF. Table 2 summarizes the information relevant to the independent variables used in the study.

Table 2

Definition of independent variables

Variable	Description
CATA	Current assets / Total assets
COMENDIV	Debt Breakdown: Current Liabilities / (Current Liabilities + Long-term liabilities)
ENDV	Current Liabilities + Non-Current Liabilities / Total Assets
FATA	Permanent assets / Total assets
GR/Sales	Gross Revenue / Sales
GRWTA	Total active growth rate
LIQCOR	Current Assets / Current Liabilities
LIQ.IMED	Available / Current Liabilities
LPA	Earnings per share
MARGL	Net Margin (%)
Net.P/CAS	Net Income / Total Current Assets
Net.P/EQTY	Net Income / book equity
Net.P/FIXA	Net Income / Permanent Assets
Net.P/TOTA	Net Income / Total Assets
PAS/PATL	Liabilities / book equity
PATL/ATNCIR	book equity/ non-current assets
PMREC	Average Receipt Term (days)

Source: Sung et al. (1999); Beaver, McNichols & Rhie (2005); Sun, Jia & Li (2011); Li, Sun & Sun (2009); Li, Sun & Wu (2010b); Li & Sun (2010); Li, Huang, Sun & Lin (2010a); Li, Lee, Zhou & Sun (2011); Li & Sun (2011); Sun et al. (2011); Li & Sun (2012); Sun et al. (2017); Wang, Veugelers & Stephan (2017).



Gujarati & Porter (2011) highlight that in data analysis, one must understand the statistical dependence between variables, avoiding non-functional or deterministic relationships. The XGBoost algorithm is an algorithm based on decision trees that uses the gradient method to build them for improving their weights. The core of the algorithm is to optimize the value of the objective function (Chen & Guestrin, 2016).

Unlike the use of coefficients to calculate the forecasting capacity that each variable has, the descending gradient method builds sequential trees to obtain the scores effectively, indicating the importance of each characteristic for the training model. The more a variable is used in the creation of the trees, the greater its weight. The algorithm considers the importance by gain, frequency and coverage (Friedman, Hastie & Tibshirani, 2001). The gain is the main reference factor for the importance of a variable in the branches of the formed trees.

3.2 SMOTE for balancing the data

The performances of machine learning algorithms are typically assessed using precision metrics. However, these may be inappropriate when the data is out of balance (Chawla, Bowyer, Hall & Kegelmeyer, 2002). The literature on the subject addresses the issue of class disproportion in two ways. One is to attribute different costs to the training sample (Pazzani, Merz, Murphy, Ali, Hume & Brunk, 1994; Domingos, 1999). The other is to produce a sample from the original data set, either by oversampling or undersampling of the majority class (Japkowicz, 2000).

In the final sample of companies analyzed over the entire period, the results show that 92% of the firm-years can be considered as solvent cases and the remainder data were defined as FD. Thus, the final sample was balanced using the method proposed by Chawla et al. (2002), in which new synthetic observations are created based on existing minority observations. We reached a balanced data after applying SMOTE, resulting in a balanced database with 52% of observations classified as non-FD and 48% of the data labelled as FD.

3.3 Techniques

Random Forest is a method based on randomly selected decision trees that employs bagging to group and obtain diverse subsets of the entire training data set and to build individual trees. The algorithm is a classification technique that consists of a selection of classifiers structured on trees $h(x, \Theta_k), k = 1, \dots, K$, where Θ_k are independent distributed random vectors and each tree casts a unit vote for the other classes of variables x . This algorithm adds extra randomness to the model, because, instead of looking for the best characteristic of the data when partitioning the nodes, it searches for the best characteristic in a random subset of the characteristics (Breiman, 1996).



In the case of XGBoost, the idea of your learning is to determine a script, where y is the expected prediction and x are the characteristic vectors (Xia et al., 2017). To build the map, the proposed model requires multiple parameters that must be predefined and the control of the appropriate combination of parameters is essential to optimize and improve the model. Below, present the parameters required in this technique (Carmona et al., 2019):

- Rounds or maximum number of iterations: the number of rounds or trees required in the model;
- Depth or maximum size of a tree: it is the number of divisions in each tree. It is used to control overfitting, as a greater depth allows the model to learn from patterns that are characteristic of a specific sample;
- Learning rate: initially introduced by Friedman (2002), it consists of a positive number (ranging from 0 to 1) that determines how quickly the algorithm adapts or the contribution of each tree to the model. A low value means that the model is more robust to overfitting;
- Gamma: minimum loss reduction required to make the next split into a node in each tree. The higher its value, the more conservative the algorithm is;
- Observation column and sample: deals with the proportion of the sub sample of variables and observations when building each tree. The observation column and sample denotes the fraction of variables and observations, which must be randomly sampled for each tree. Its value ranges from 0 to 1, avoids overfitting and speeds up the algorithm's calculations;
- Minimum child weight: indicates the minimum sum of the necessary instance weight in a node. If the tree separation step results in a leaf node whose sum of previous weight is less than the value assigned to that parameter, the construction process will interrupt further subdivisions.

Regularization term or penalty on weights: the regularization term controls the complexity of the model that helps to avoid over-adjustment.

Logistic regression is one of the most accepted and used techniques in theoretical terms, since two distinct classes (good or bad) are defined (Kim, 2011; Li et al., 2011; Li & Sun,

2011). Given a set of N training records $D = (x_i, y_i) \ i = 1, \dots, N$, with the independent variables $x_i \in \mathbb{R}$ and dependent matching binary variables $y_i \in \{0, 1\}$, the logistic regression estimates the probability of FD, or $P(y=1|x)$, as follows:

$$P(y = 1|x) = \frac{\exp((\beta_i^T x_i))}{1 + \exp((\beta_i^T x_i))} \quad (1)$$

where β_i are the regression coefficients estimated by maximum likelihood.

In terms of techniques, we compare each performance with one another, aiming to find the best one considering some measures of prediction power. We define these measures in the following section.

3.4 Performance measures

We evaluate five traditional assessment metrics, which are overall accuracy (ACC), type I and type II errors (T1E and T2E), the area under the ROC curve (AUC) and the Kolmogorov-Smirnov statistic (KS) to assess the performance of models. The main assessment metric is AUC (Area under the curve), which is an alternative measure of discrimination capacity based on the ROC curve. The ROC curve shows true positive rate (TPR) values versus false positive rate (FPR) values in various limit settings. The true positive rate is also known as sensitivity and the false positive rate is known as specificity and can be calculated as follows ($1 - \text{specificity}$). In other words, the AUC score measures how well the model discriminates between the two classes (Wang et al., 2017; Zhang, Priestley & Ni, 2018).

Regarding the XGBoost calibration, after adjusting the parameters according to Carmona et al. (2019), we achieved the best model based on the following values: number of iterations: 1,000; maximum depth: 5; learning rate: 0.1; range: 0; observation sample: 0.8; minimum child weight: 1; regularization: 0. The intention of calibrating and controlling the parameters is to avoid overfitting the model to the data and ensuring its generalization.

The performance of each model is evaluated on a different set of data than the one used to train it. Thus, we split the dataset into 70% for training and 30% for testing the models. In the larger fragment of data, each model was trained and adjusted, while the smaller fragment, used to test the model, following (Li, Tian, Li, Zhou & Yang, 2017; Xia et al., 2018).

The confusion matrix contains four basic elements: true positives (TP) indicating that the default forecast is consistent with its real value; false negatives (FN) means that the result of the sample forecast is classified as non-performing, but its actual label indicates non-performing. Likewise, false positives (FP) are those samples of FD firms classified as FD and those within samples of FD correctly predicted to be FD are labeled as true negatives (TN). All these values are inputs for evaluating the ACC, the errors.

The ACC, given by Equation 2, is the model's precision measure compared to the general data. It is the ratio between the units correctly classified and the total number of forecasts made by the classifiers. ACC is calculated as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$



The Under the ROC Curve (AUC) measures the ability of a binary model, in other words, a way to represent the ROC curve in single value, adding all ROC thresholds calculating the "area under the curve". The limit for two classes is 0.5 and the closer to the value 1, the better the algorithm is able to distinguish between classes.

Use the Kolmogorov-Smirnov (KS) statistic (Equation 3) to measure the adequacy of the model, where TPR stands for the rate of true positives and FRP is the rate of false positives. KS is one of the most useful and widespread non-parametric methods for comparing FD and non-FD firms correctly predicted (Zhang et al., 2018).

$$KS = \max_t (|TPR(t) - FPR(t)|) \quad (3)$$

Finally, we use type I and type II error rates as indicators to further explore the model's ability to classify customers as non-performing or non-performing, respectively. In this experiment, the Type I error rate (T1E) denotes the proportion of misclassified non-FD firms, and the Type II error rate (T2E) refers to the proportion of misclassified FD instances. Their values are calculated according to Equations 4 and 5:

$$T1E = \frac{TN}{FP + TN} \quad (4)$$

$$T2E = \frac{TP}{FN + TP} \quad (5)$$

4 Results and discussions

Table 3 shows the descriptive statistics of the study, showing information relevant to the number of observations. It is possible to observe that the average LPA in the period is negative, therefore, it can be inferred that the companies were operating with low margins, accumulating losses, in some way. The most widely dispersed variable is NETPEQTY, which deals with the company's ability to add value from its own resources and from investors, which suggests that there are companies that are losing more value in each investment made and others that add more. The least dispersed variable is COMENDIV, which deals with the composition of indebtedness, which shows that companies have, on average, short-term debt, mainly because long-term financing and lower costs are less accessible.

Table 3

Descriptive statistics of the raw data (i.e., before SMOTE), these data came from 17,282 instances (firm-year)

Variable	Mean	SD	Minimum	Maximum
PMREC	229.290	2.998.601	-8.423.438	326.606.000
ENDV	1.114	7.402	0.000	487.224
LIQIMED	0.829	3.319	-1.000	159.257
CATA	0.373	0.204	0.000	1.000
COMENDIV	0.728	0.274	0.000	1.000
LIQCOR	1.831	3.345	0.000	160.257
LPA	-5.539	84.377	-2.364.538	2.661.399
PASPATL	1.447	24.590	-901.545	1.148.586
PATLATNCIR*	130bi	8.9bi	-753.173	860bi
FATA	0.548	0.275	0.000	1.000
GRSALES	5.377.712	706.899.700	-3.948	9.29E+07
GRWTA	7.204	848.929	-0.964	111.083.200
NETPCAS	-1.331	29.306	-1.207.017	618.788
NETPEQTY	-0.004	8.458	-1.066.615	109.000
NETPFIXA*	18.7mi	1.61bi	-25.6bi	188bi
NETPTOTA	-0.047	0.689	-46.334	11.889
MARGL	-1.445.015	33.932.040	-1.899.698.000	132.037.500
MARKVAL*	1.65bi	7.11bi	0.000	243bi

Source: Research data. Note: SD means Standard Deviation; * bi = billions and mi = millions.

Since correlated variables can hinder the model's performance, we observed the final dataset for checking multicollinearity problems (Carmona et al., 2019). Table 4 shows the correlation matrix of all variables tested in our work. Before carrying out the analyzes, no correlated variable was found. In particular, the biggest association involving the dependent variable (FINDISTR) occurred with FATA (the ratio of fixed assets to total assets, that is, how much capital is invested in assets that helps to increase revenue).

Table 4

Pearson's correlation matrix

Variable	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]	[15]	[16]	[17]	[18]
[1]PMREC	-																	
[2]ENDV	0.00	-																
[3]LIQIMED	-0.01	-0.03	-															
[4]CATA	0.06	-0.05	0.16	-														
[5]LIQCOR	-0.01	-0.03	0.95	0.15	-													
[6]COMENDIV	0.00	-0.08	-0.02	0.10	-0.02	-												
[7]LPA	0.00	-0.01	0.03	0.03	0.03	0.01	-											
[8]PASPATL	0.00	0.00	-0.03	0.00	-0.03	-0.04	0.01	-										
[9]PATLATNCIR	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	-									
[10]FATA	-0.03	0.04	-0.12	-0.57	-0.13	-0.27	-0.02	0.02	-0.03	-								
[11]GRSALES	0.00	0.00	0.00	0.01	0.00	-0.01	0.00	0.00	0.00	0.00	-							
[12]GRWTA	0.00	0.00	0.01	0.02	0.01	0.00	0.00	0.00	0.00	-0.01	0.00	-						
[13]NETPCAS	0.00	-0.16	0.02	0.04	0.02	0.03	0.02	0.00	0.00	-0.04	0.00	0.00	-					
[14]NETPEQTY	0.00	0.00	0.01	0.02	0.01	0.00	0.02	-0.42	0.00	-0.02	0.00	0.00	0.02	-				
[15]NETPFIXA	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.96	-0.02	0.00	0.00	0.00	0.00	-			
[16]NETPTOTA	0.00	-0.79	0.03	0.04	0.03	0.06	0.05	-0.01	0.00	-0.03	0.00	0.00	0.18	0.04	0.00	-		
[17]MARGL	0.00	-0.34	-0.03	0.05	-0.03	0.06	0.00	0.00	0.00	-0.02	0.00	0.00	0.23	0.00	0.00	0.39	-	
[18]MARKVAL	-0.01	-0.02	-0.01	-0.05	-0.01	-0.06	0.02	0.00	0.00	0.07	0.00	0.00	0.01	0.00	0.00	0.04	0.01	-
FINDISTR	0.03	0.06	-0.06	-0.04	-0.06	0.07	-0.07	-0.03	-0.01	0.17	-0.01	-0.01	-0.04	0.01	-0.01	-0.08	-0.01	-0.06

Source: Research data.

Table 5 shows the performance measures of the models tested in this research, comparing them with the results of Carmona et al. (2019) who tested the same models, although applied to the banking sector. Our results followed the same trend of AUC of the models presented by the authors, where the XGBoost was what got the best result, followed by RF and finally the logistic regression.


Table 5

Performance measures for the models developed in this work (our results), comparing to Carmona et al.

(2019) (Benchmark)

Model	Benchmark*		Our Results				
	ACC	AUC	ACC	AUC	KS	T1E	T2E
XGBoost	0.9474	0.98	0.9605	0.9636	0.3964	0.0471	0.0312
Random Forest	0.9211	0.93	0.9510	0.9502	0.3905	0.0411	0.0571
Logistic Regression	0.8421	0.84	0.6512	0.6751	0.1439	0.3108	0.3782

Source: research data. * Retrieved from Carmona et al. (2019). ACC, AUC, KS, T1E and T2E are overall accuracy, area under ROC curve, Kolmogorov-Smirnov Statistic, Type I and Type II errors, respectively.



The XGBoost model is particularly useful for solving classification problems, as it minimizes general errors when generating models based on errors in previous tree iterations. For the Latin American market, using machine learning models as the basis for this study, it obtained better results than Logistic Regression. The XGBoost had a slightly better result than the RF either, about 0.0134 percentage points higher AUC. Following the same direction as the studies by Barboza et al. (2017), where the Boosting (0.9297) and RF (0.9292) techniques were highly accurate, but as in our studies, these models had little variation between their results. Our results are still in line with the results presented by Chang & Hsu (2018), who despite evaluating a market different from the banking sector, the XGBoost model (0.94) stood out in comparison with the LR models (0.77), SVM (0.87) and Neural Networks (0.82). Yet, our outcomes are close to Zięba et al. (2016) in which the XGBoost get the best result in the first year of forecasts (0.959).

Sun et al. (2017) used the AdaBoost support vector machine model in their study, an algorithm that relies on errors from previous classifications to make the next classification also sequentially, such as XGBoost. In this study, the authors applied the model to predict situations of FD for Chinese companies and achieved an accuracy of 93.88%. In a similar study, Kim et al. (2015) use GMboost (Geometric Mean based Boosting) to predict bankruptcy in Korean companies and the model reaches an accuracy of 82%.

Korol (2013) also investigated and compared statistical methods with computational models in Latin American companies, the author argues that the structure of the models that predict bankruptcy in these companies is much more complicated and complex than the models that predict failure in European companies. The author states that although there is a consensus in the literature on the most popular credit forecasting systems, each country or region must develop its own, due to the different accounting systems and different attributes of economies between countries, each country or region in the world. It must have its own bankruptcy forecasting model, which takes into account the structure of the domestic financial system. In line with the author and also with the works of Barboza et al. (2017), our research contributes to the discussion of the techniques based on traditional statistics and the the computational methods, where they obtained better results.

Thus, we could infer that the AUC result of 0.9636 presented in our model is satisfactory given the complexity of this emerging market, which is Latin America and yet the variation in relation to the results of the machine learning models presented in particular by Carmona et al. (2019), demonstrating the consistency of the XGBoost and RF models. On the contrary, the Logistic Regression had a much lower result, around 16.49%. In this case, a suggestion would be to include other variables that could better adhere to the model.

As for the value of the KS statistic, the XGBoost model presented a slightly better result, 39.64%, thus proving its promising ability to discriminate between two types of customer types.

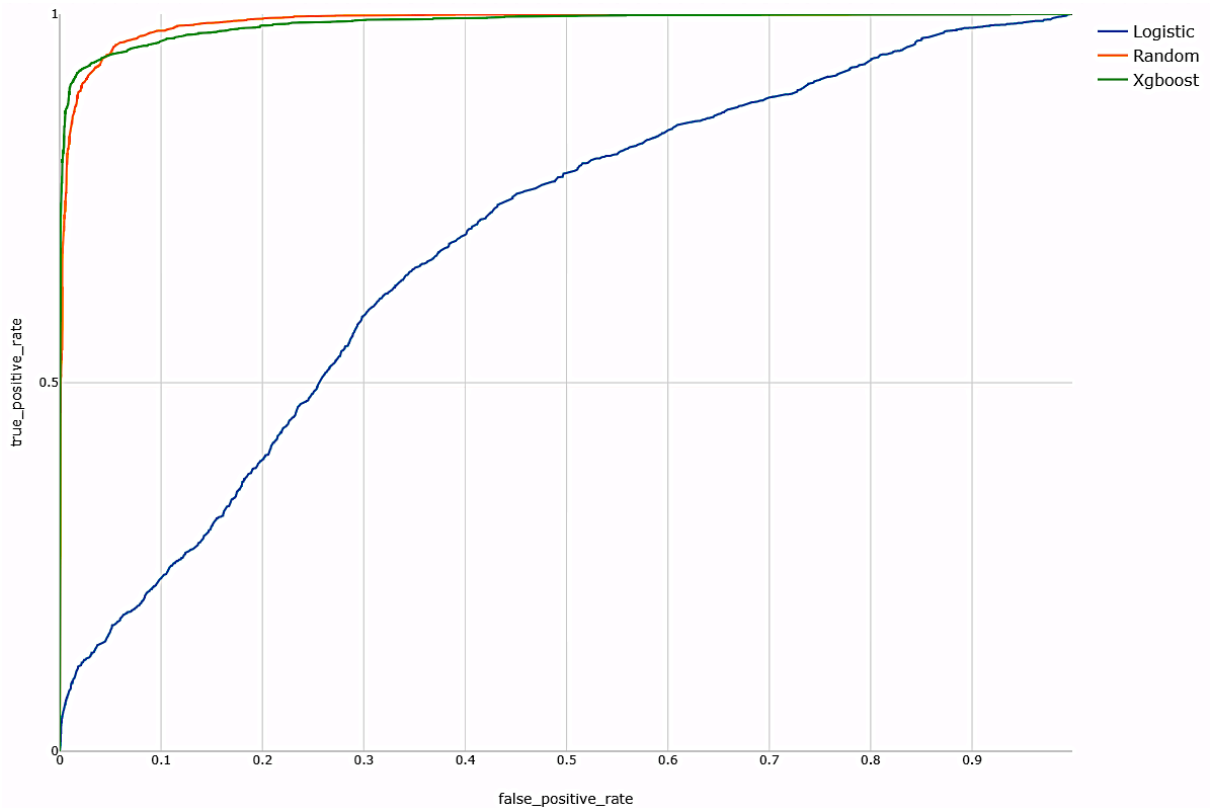


Type I error is related to the non-forecasting of a FD company, which leads to recovery costs, financial losses, opportunity costs, sunk interest, being more costly than the type II error, which is related to a company that does not present this picture, however it is labeled as such (West, 2000; Abdou, 2009). Evaluating the model presented in this research, can see that the XGBoost model obtained the best error rates, compared with the other models, in together with RF, which reinforces model consistency and machine learning. In the study conducted by Barboza et al. (2017) the errors were greater when compared to our studies, with Boosting (T1E = 18.8% and T2E = 13.3%) and RF (T1E = 16.5% and T2E = 12.9%), which suggests assertiveness in the selection of variables used for the Latin American market.

The area under the ROC curve (AUC) is an important indicator, as it provides us with an independent measure of overall accuracy. The value of the area below the diagonal (0.5 or 50%) is not valid, as they are considered as random. However, a value that approaches 1.0 (or 100%) shows best ability of the model for making correct predictions. It also shows in an illustrative way the capacity for discernment between classes. Figure 1 shows the comparison of the three models. The XGBoost model reveals a satisfactory performance but, when compared to the Random Forest model the difference is slightly inferior to the proposed model. In general, the machine learning models had better accuracy than the Logistic Regression model, which demonstrates the complexity of predicting FD in Latin American firms.

Figure 1

ROC curve for each model tested



Source: Research data.

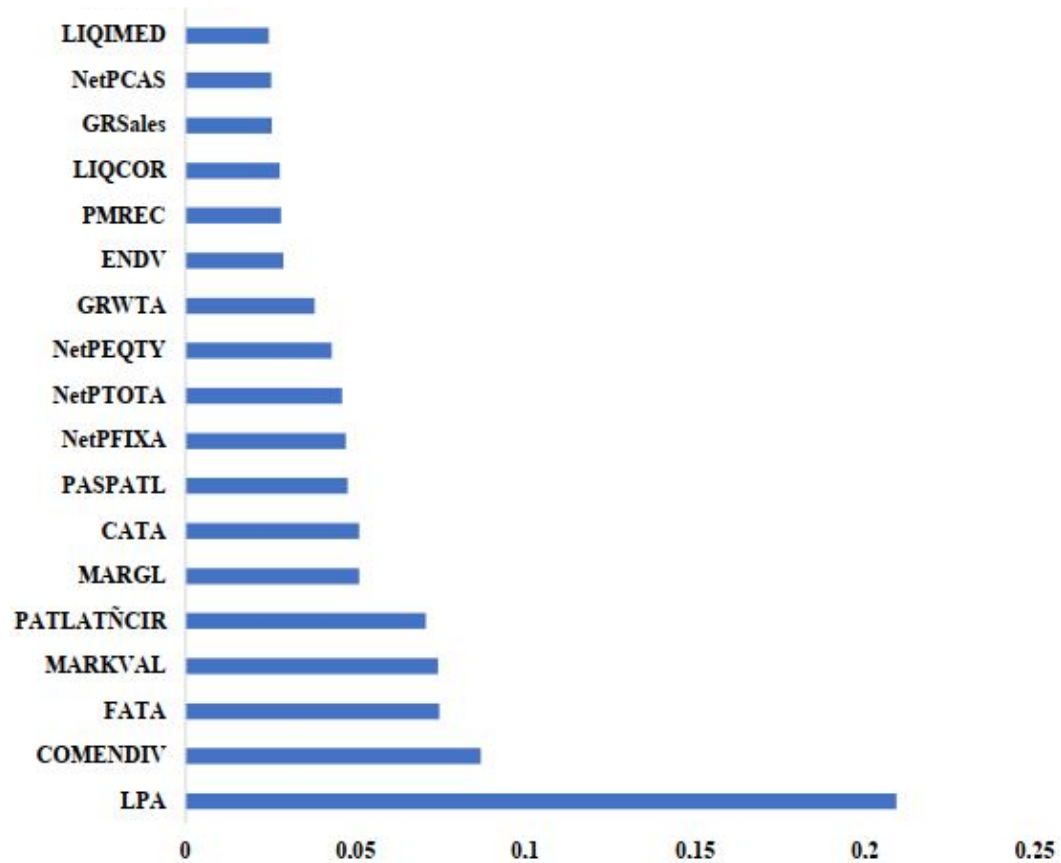
Another interesting aspect of XGBoost is that its results expose the variables with the greatest influence on the prediction of the dependent variable, thus corroborating with Abdou (2009). In his study, the researcher suggests future investigations about which variables may indicate possible financial distress in advance. This becomes an important issue because companies can track and monitor specific characteristics of their operations in order to avoid financial inconvenience.

In Figure 2, the most relevant variables with the greatest relative influence on the response variable are listed: LPA, COMENDIV, FATA and MARKVAL. We can observe that the variables LPA and COMENDIV provided the highest discriminatory values. From this result, it is possible to imply an alert that companies already in FD would be committed to retaining funds for the payment of interest and debt amortization, which would trigger the correlation between indebtedness and non-payment of dividends.



Figure 2

Variables of greater weight identified by the XGBoost model




Source: Research data.

Thus, can be highlighted that profitability is one of the main sources of business financing, that is, the company remains liquid at the disposal of its own operation. Although the maintenance of shareholders with the payment of dividends is relevant to the business, maintaining the operation of the firm and its financial liquidity brings better long-term results. Thus, dividend payment policies should be prioritized in view of market and commercial conditions that the company is experiencing (Sun et al., 2011; Barboza et al., 2017; Carmona et al., 2019). Zieba et al. (2016) also found relevance in profitability and leverage for predicting bankruptcies.

The composition of indebtedness was also relevant in forecasting FD as deals with the fundraising policy within organizations, and reveals the strategy of the companies in relation to their short-term obligations. Sun et al. (2011) and Huang & Yen (2019) also included this variable in their models for forecasting bankruptcy and FD.

Investment projects and payment of dividends compete for the same sources of funds, so the variable FATA – which is also known as asset turnover – reflects the efficiency with which the company uses its assets, that is, it specifically measures a company's ability to generate sales from investment in



fixed assets. In general, a balanced asset turnover rate indicates that a company has efficiently used investment in fixed assets in its operation to generate revenue.

Sung et al. (1999) also found relevance for this variable, and characterized this variable as a key to evaluate companies in FD. The authors claim that this proportion refers to the company's reliability or security in fixed asset financing, if this index is too high, the company is suffering from a shortage of current assets and will not be able to pay short-term loans, but if this proportion is very low, the company is not using its capital efficiently, as there is an excess of investments in fixed assets.

The most valued companies in the capital market (MARKVAL) are less likely to get into financial distress and this variable was relevant to the XGBoost model, as well as in the studies by Barboza et al. (2017). Beaver et al. (2005) argue that the market value variable absorbs much of the predictive power of the indicators in the financial statements and provides additional explanatory power not reflected in the financial indexes.

Our results were different in relation to the relevance of the variables in the studies by Carmona et al. (2019) that used the same models, but in a different market. The variables that highlighted the greatest weight in the model presented by the authors were: the return before taxes on assets, the total capital ratio based on risk, the accumulated profits on average equity and the return on profitable assets. However, our results reveal interesting ability of predicting FD.

In view of the results presented, we argue that the predictive precision of machine learning models is considerably better, to deal with more complex markets, as in Latin America, multiple variables, extensive, non-linear and misadjusted databases and it is still easy to apply, when compared with traditional statistical models (Korol, 2013; Barboza et al., 2017). Considering the XGBoost technique, the regularization of the gradient increase for weight control and selection of variables, considerably improved the predictive precision in relation to the other models, in this way, we can deduce that the XGBoost also contributes to the reduction of overfitting, but it is necessary to moderation of the appropriate combination of parameters to maximize the model, and consequently improve accuracy (Carmona et al. (2019).

5 Concluding remarks

The main objective of this research was to predict the probability of a company going into a financial distress situation by using three different techniques: Extreme Gradient Boosting (XGBoost), Random Forest, both as machine learning instances, and Logistic Regression, as a benchmark, widely used in the literature. XGboost is an evolution of other methods - such as AdaBoost and Random Forest - and has been applied in recent studies in order to predict bank failures and credit scores.



As research contributions allow the expansion of knowledge through the application of a financial distress prediction model from the perspective of machine learning techniques in publicly traded companies in Latin America, its application in this database is unprecedented. The study further expands the discussion of the topic by focusing on the predictive power of model comparison, highlighting the benefits of using machine learning algorithms applied to financial research.

As a practical contribution of this research, it is believed that through analysis, companies can develop preventive systems that alert their managers about indicators that may affect their financial health. In addition, financial institutions could avoid default by taking appropriate precautionary measures, instead of waiting until financial constraint occurs. The forecast of financial distresses also serves as an aid in the decision of external investors and creditors, being an effective form of risk management.

As a social contribution it allows us to glimpse that when companies are unable to pay their obligations due to financial distress, the chances of banks not receiving their money back increase and this would cause problems for the entire financial system. So efficient forecasting models allow both companies to improve their risk management and banks to make more assertive credit decisions, which is crucial for a healthy financial system in a society.

The study showed that XGBoost has greater predictive power when compared to the others, considering the parameters used here. Furthermore, this method has an important feature, which corroborates with Chen et al. (2011), Wang & Ma (2011), Tsai et al. (2014), Zhao, Xu, Kang, Kabir, Liu & Wasinger (2015), where they suggest that models and results should focus on explaining the reasons why companies faced financial distress; factors that are important for both corporations and financial institutions. As there is no concrete answer on the most representative characteristics (independent variables), the proposed method demonstrated that the reduction in the distribution of dividends classifies a company with a high probability of facing a future financial tightening and its investments also contribute to this.

The predictive ability of the model tested in this study should encourage further research to join forces with computer scientists and add dynamics to the econometric models commonly adopted in finance studies. In addition, in an attempt to extend the current limits of performance and interpretability, XGBoost tracks variables that can add extra predictive weight to the model and bring operational agility to the entire process.

As a limitation of the research, we stand out that parameters should be further explored and other techniques, such as Neural Networks, could counter our results. However, even in the face of the limitation arising from possible endogeneity of the model, the study moves forward by reinforcing and stimulating the automation of the processes and models in this context. On the one hand, a large volume of different data can be extremely beneficial for machine learning; on the other hand, is necessary to

think about celerity in the decision-making process and this will be achieved with the greatest number of relevant information about customers, developments of low-cost algorithms and that maximize the operational capacity of the machines.

In view of the results, it is believed that companies can develop preventive systems that would alert their managers about indicators that may affect their financial health. In addition, institutions could avoid default by taking appropriate precautionary measures, rather than waiting until the financial constraint takes place.

In view of the facts and results found in this research, we suggest that future studies of this article are undoubtedly of interest in improving the model, observing some aspects, among which this necessary article is related:

- Encourage the construction and expansion of databases with different variables, however, relevant to credit research that could improve the robustness of the model;
- Researchers should structure databases that contain diverse and relevant variables in the area - both financial information and qualitative aspects - and encourage their adoption in financial institutions to improve, thus, both the process and the forecasts of new models;
- Assess the ideal division of a database for training and testing the model;
- Experiment with different parameters on different data, such as corporate variables or information from companies in emerging markets;
- Include qualitative variables (for example, social, macroeconomic information, quality of management, among others).

References

- Abdou, H. A. (2009). Genetic programming for credit scoring: The case of Egyptian public sector banks. *Expert Systems with Applications*, 36, 11402–11417. <https://doi.org/10.1016/j.eswa.2009.01.076>
- Alfaro, E., García, N., Gámez, M., & Elizondo, D. (2008). Bankruptcy forecasting: An empirical comparison of adaboost and neural networks. *Decision Support Systems*, 45, 110–122. <https://doi.org/10.1016/j.dss.2007.12.002>
- Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23, 589–609. <https://doi.org/10.1111/j.1540-6261.1968.tb00843.x>
- Bae, J. K. (2012). Predicting financial distress of the South Korean manufacturing industries. *Expert Systems with Applications*, 39, 9159–9165. <https://doi.org/10.1016/j.eswa.2012.02.058>
- 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>



- Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of accounting research*, (pp. 71–111). <https://www.jstor.org/stable/2490171>
- Beaver, W. H., McNichols, M. F., & Rhie, J.-W. (2005). Have financial statements become less informative? evidence from the ability of financial ratios to predict bankruptcy. *Review of Accounting studies*, 10, 93–122. <https://doi.org/10.1007/s11142-004-6341-9>
- BIS (2006). *Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework – Comprehensive Version*. Retrieved from: <http://www.bis.org/publ/bcbs128.pdf>
- Blanco, A., Pino-Mejías, R., Lara, J., & Rayo, S. (2013). Credit scoring models for the microfinance industry using neural networks: Evidence from peru. *Expert Systems with applications*, 40, 356–364. <https://doi.org/10.1016/j.eswa.2012.07.051>
- Breiman, L. (1996). Bagging predictors. *Machine learning*, 24, 123–140. <https://doi.org/10.1007/BF00058655>
- Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 39, 3446–3453. <https://doi.org/10.1016/j.eswa.2011.09.033>
- Campos, A. L. S., & Nakamura, W. T. (2015). Rebalanceamento da estrutura de capital: endividamento setorial e folga financeira. *Revista de Administração Contemporânea*, 19, 20–37. <https://doi.org/10.1590/1982-7849rac20151789>
- 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>
- Chang, T.-M., & Hsu, M.-F. (2018). Integration of incremental filter-wrapper selection strategy with artificial intelligence for enterprise risk management. *International Journal of Machine Learning and Cybernetics*, 9, 477–489. <https://doi.org/10.1007/s13042-016-0545-8>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- Chen, H.-L., Yang, B., Wang, G., Liu, J., Xu, X., Wang, S.-J., & Liu, D.-Y. (2011). A novel bankruptcy prediction model based on an adaptive fuzzy k-nearest neighbor method. *Knowledge-Based Systems*, 24, 1348–1359. <https://doi.org/10.1016/j.knsys.2011.06.008>
- Chen, T., & Guestrin, C. (2015). XGBoost: reliable large-scale tree boosting system. In *Proceedings of the 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA* (pp. 13–17). Available: http://ml-pai-learn.oss-cn-beijing.aliyuncs.com/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0%E8%B5%84%E6%96%99/LearningSys_2015_paper_32.pdf
- Cho, S., Hong, H., & Ha, B.-C. (2010). A hybrid approach based on the combination of variable selection using decision trees and case-based reasoning using the mahalanobis distance: For bankruptcy

prediction. *Expert Systems with Applications*, 37, 3482–3488.
<https://doi.org/10.1016/j.eswa.2009.10.040>

Chuang, C.-L., & Huang, S.-T. (2011). A hybrid neural network approach for credit scoring. *Expert Systems*, 28, 185–196. <https://doi.org/10.1111/j.1468-0394.2010.00565.x>

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>

Danenas, P., & Garsva, G. (2015). Selection of support vector machines based classifiers for credit risk domain. *Expert Systems with Applications*, 42, 3194–3204.
<https://doi.org/10.1016/j.eswa.2014.12.001>

Domingos, P. (1999). Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 155–164). <https://dl.acm.org/doi/pdf/10.1145/312129.312220>

Finlay, S. (2011). Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research*, 210, 368–378.
<https://doi.org/10.1016/j.ejor.2010.09.029>

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. Springer series in statistics New York.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, (pp. 1189–1232). <https://www.jstor.org/stable/2699986>

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, 38, 367–378. [https://doi.org/10.1016/S0167-9473\(01\)00065-2](https://doi.org/10.1016/S0167-9473(01)00065-2)

Gepp, A., Kumar, K., & Bhattacharya, S. (2010). Business failure prediction using decision trees. *Journal of forecasting*, 29, 536–555. <https://doi.org/10.1002/for.1153>

Gujarati, D. N., & Porter, D. C. (2011). *Econometria Básica-5*. Amgh Editora.

Guo, Y., Zhou, W., Luo, C., Liu, C., & Xiong, H. (2016). Instance-based credit risk assessment for investment decisions in p2p lending. *European Journal of Operational Research*, 249, 417–426.
<https://doi.org/10.1016/j.ejor.2015.05.050>

Harris, T. (2015). Credit scoring using the clustered support vector machine. *Expert Systems with Applications*, 42, 741–750. <https://doi.org/10.1016/j.eswa.2014.08.029>

He, H., Zhang, W., & Zhang, S. (2018). A novel ensemble method for credit scoring: Adaption of different imbalance ratios. *Expert Systems with Applications*, 98, 105–117.
<https://doi.org/10.1016/j.eswa.2018.01.012>

Hens, A. B., & Tiwari, M. K. (2012). Computational time reduction for credit scoring: An integrated approach based on support vector machine and stratified sampling method. *Expert Systems with Applications*, 39, 6774–6781. <https://doi.org/10.1016/j.eswa.2011.12.057>

Heo, J., & Yang, J. Y. (2014). Adaboost based bankruptcy forecasting of korean construction companies. *Applied soft computing*, 24, 494–499. <https://doi.org/10.1016/j.asoc.2014.08.009>




- Hsieh, T.-J., Hsiao, H.-F., & Yeh, W.-C. (2012). Mining financial distress trend data using penalty guided support vector machines based on hybrid of particle swarm optimization and artificial bee colony algorithm. *Neurocomputing*, 82, 196–206. <https://doi.org/10.1016/j.neucom.2011.11.020>
- Huang, Y.-P., & Yen, M.-F. (2019). A new perspective of performance comparison among machine learning algorithms for financial distress prediction. *Applied Soft Computing*, 83, 105663. <https://doi.org/10.1016/j.asoc.2019.105663>
- Iquipaza, R. A., Lamounier, W. M., & Amaral, H. F. (2008). Assimetric information and dividends payout at the São Paulo stock exchange (bovespa). *Ad. Sci. appl. Account.* 1(1), 1001-14. Available: https://www.researchgate.net/publication/241765395_Asymmetric_information_and_dividends_payout_at_the_Sao_Paulo_stock_exchange
- Iturriaga, F. J. L., & Sanz, I. P. (2015). Bankruptcy visualization and prediction using neural networks: A study of us commercial banks. *Expert Systems with applications*, 42, 2857– 2869. <https://doi.org/10.1016/j.eswa.2014.11.025>
- Japkowicz, N. (2000). The class imbalance problem: Significance and strategies. In *Proc. of the Int'l Conf. on Artificial Intelligence*. Available: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.35.1693&rep=rep1&type=pdf>
- Jeong, C., Min, J. H., & Kim, M. S. (2012). A tuning method for the architecture of neural network models incorporating gam and ga as applied to bankruptcy prediction. *Expert Systems with Applications*, 39, 3650–3658. <https://doi.org/10.1016/j.eswa.2011.09.056>
- Jones, S., Johnstone, D., & Wilson, R. (2015). An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes. *Journal of Banking & Finance*, 56, 72–85. <https://doi.org/10.1016/j.jbankfin.2015.02.006>
- Kim, M.-J., & Kang, D.-K. (2010). Ensemble with neural networks for bankruptcy prediction. *Expert systems with applications*, 37, 3373–3379. <https://doi.org/10.1016/j.eswa.2009.10.012>
- Kim, M.-J., Kang, D.-K., & Kim, H. B. (2015). Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction. *Expert Systems with Applications*, 42, 1074–1082. <https://doi.org/10.1016/j.eswa.2014.08.025>
- Kim, S. Y. (2011). Prediction of hotel bankruptcy using support vector machine, artificial neural network, logistic regression, and multivariate discriminant analysis. *The Service Industries Journal*, 31, 441–468. <https://doi.org/10.1016/j.eswa.2014.08.025>
- Kim, S. Y., & Upneja, A. (2014). Predicting restaurant financial distress using decision tree and adaboosted decision tree models. *Economic Modelling*, 36, 354–362. <https://doi.org/10.1016/j.econmod.2013.10.005>
- Korol, T. (2013). Early warning models against bankruptcy risk for central european and latin american enterprises. *Economic Modelling*, 31, 22–30. <https://doi.org/10.1016/j.econmod.2012.11.017>
- Lai, K. K., Yu, L., Wang, S., & Zhou, L. (2006). Credit risk analysis using a reliability-based neural network ensemble model. In *International Conference on Artificial Neural Networks* (pp. 682–690). Springer, Berlin, Heidelberg. https://doi.org/10.1007/11840930_71

- 
- Lee, T.-S., Chiu, C.-C., Lu, C.-J., & Chen, I.-F. (2002). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with applications*, 23, 245–254. [https://doi.org/10.1016/S0957-4174\(02\)00044-1](https://doi.org/10.1016/S0957-4174(02)00044-1)
- Li, H., Huang, H.-B., Sun, J., & Lin, C. (2010a). On sensitivity of case-based reasoning to optimal feature subsets in business failure prediction. *Expert Systems with Applications*, 37, 4811–4821. <https://doi.org/10.1016/j.eswa.2009.12.034>
- Li, H., Lee, Y.-C., Zhou, Y.-C., & Sun, J. (2011). The random subspace binary logit (rsbl) model for bankruptcy prediction. *Knowledge-Based Systems*, 24, 1380–1388. <https://doi.org/10.1016/j.knosys.2011.06.015>
- Li, H., & Sun, J. (2010). Business failure prediction using hybrid2 case-based reasoning (h2cbr). *Computers & Operations Research*, 37, 137–151. <https://doi.org/10.1016/j.cor.2009.04.003>
- Li, H., & Sun, J. (2011). Empirical research of hybridizing principal component analysis with multivariate discriminant analysis and logistic regression for business failure prediction. *Expert Systems with Applications*, 38, 6244–6253. <https://doi.org/10.1016/j.eswa.2010.11.043>
- Li, H., & Sun, J. (2012). Forecasting business failure: The use of nearest-neighbour support vectors and correcting imbalanced samples—evidence from the chinese hotel industry. *Tourism Management*, 33, 622–634. <https://doi.org/10.1016/j.tourman.2011.07.004>
- Li, H., Sun, J., & Sun, B.-L. (2009). Financial distress prediction based on or-cbr in the principle of k-nearest neighbors. *Expert Systems with Applications*, 36, 643–659. <https://doi.org/10.1016/j.eswa.2007.09.038>
- Li, H., Sun, J., & Wu, J. (2010b). Predicting business failure using classification and regression tree: An empirical comparison with popular classical statistical methods and top classification mining methods. *Expert Systems with Applications*, 37, 5895–5904. <https://doi.org/10.1016/j.eswa.2010.02.016>
- Li, Z., Tian, Y., Li, K., Zhou, F., & Yang, W. (2017). Reject inference in credit scoring using semi-supervised support vector machines. *Expert Systems with Applications*, 74, 105–114. <https://doi.org/10.1016/j.eswa.2017.01.011>
- Malhotra, R., & Malhotra, D. K. (2002). Differentiating between good credits and bad credits using neuro-fuzzy systems. *European journal of operational research*, 136, 190–211. [https://doi.org/10.1016/S0377-2217\(01\)00052-2](https://doi.org/10.1016/S0377-2217(01)00052-2)
- Manzaneque, M., Priego, A. M., & Merino, E. (2016). Corporate governance effect on financial distress likelihood: Evidence from Spain. *Revista de Contabilidad*, 19, 111–121. <https://doi.org/10.1016/j.rcsar.2015.04.001>
- Martinez-Villa, B. A., & Machin-Mastromatteo, J. D. (2016). Four theories to improve justice in Latin America. *Information Development*, 32, 1284–1288. <https://doi.org/10.1177/0266666916658588>
- Pazzani, M., Merz, C., Murphy, P., Ali, K., Hume, T., & Brunk, C. (1994). Reducing misclassification costs. In *Machine Learning Proceedings 1994* (pp. 217–225). Elsevier, (pp. 217-225). Morgan Kaufmann. <https://doi.org/10.1016/B978-1-55860-335-6.50034-9>



- Pindado, J., Rodrigues, L., & De La Torre, C. (2008). Estimating financial distress likelihood. *Journal of Business Research*, 61, 995–1003. <https://doi.org/10.1016/j.jbusres.2007.10.006>
- Piramuthu, S. (1999). Financial credit-risk evaluation with neural and neuro fuzzy systems. *European Journal of Operational Research*, 112, 310–321. [https://doi.org/10.1016/S0377-2217\(97\)00398-6](https://doi.org/10.1016/S0377-2217(97)00398-6)
- Sanz, L. J., & Ayca, J. (2006). Financial distress costs in Latin America: A case study. *Journal of Business Research*, 59, 394–395. <https://doi.org/10.1016/j.jbusres.2005.09.014>
- Sun, J., Fujita, H., Chen, P., & Li, H. (2017). Dynamic financial distress prediction with concept drift based on time weighting combined with adaboost support vector machine ensemble. *Knowledge-Based Systems*, 120, 4–14. <https://doi.org/10.1016/j.knosys.2016.12.019>
- Sun, J., Jia, M.-Y., & Li, H. (2011). Adaboost ensemble for financial distress prediction: An empirical comparison with data from chinese listed companies. *Expert Systems with Applications*, 38, 9305–9312. <https://doi.org/10.1016/j.eswa.2011.01.042>
- Sung, T. K., Chang, N., & Lee, G. (1999). Dynamics of modeling in data mining: interpretive approach to bankruptcy prediction. *Journal of Management Information Systems*, 16, 63–85. <https://doi.org/10.1080/07421222.1999.11518234>
- Tang, X., Li, S., Tan, M., & Shi, W. (2020). Incorporating textual and management factors into financial distress prediction: A comparative study of machine learning methods. *Journal of Forecasting*. <https://doi.org/10.1002/for.2661>
- Tsai, C.-F., Hsu, Y.-F., & Yen, D. C. (2014). A comparative study of classifier ensembles for bankruptcy prediction. *Applied Soft Computing*, 24, 977–984. <https://doi.org/10.1016/j.asoc.2014.08.047>
- Tsai, C.-F., & Wu, J.-W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert systems with applications*, 34, 2639–2649. <https://doi.org/10.1016/j.eswa.2007.05.019>
- Wang, G., & Ma, J. (2011). Study of corporate credit risk prediction based on integrating boosting and random subspace. *Expert Systems with Applications*, 38, 13871–13878. <https://doi.org/10.1016/j.eswa.2011.04.191>
- Wang, G., Ma, J., & Yang, S. (2014). An improved boosting based on feature selection for corporate bankruptcy prediction. *Expert Systems with Applications*, 41, 2353–2361. <https://doi.org/10.1016/j.eswa.2013.09.033>
- Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, 46, 1416–1436. <https://doi.org/10.1016/j.respol.2017.06.006>
- Wang, Y., Wang, S., & Lai, K. K. (2005). A new fuzzy support vector machine to evaluate credit risk. *IEEE Transactions on Fuzzy Systems*, 13, 820–831. <https://doi.org/10.1109/TFUZZ.2005.859320>
- West, D. (2000). Neural network credit scoring models. *Computers & Operations Research*, 27, 1131–1152. [https://doi.org/10.1016/S0305-0548\(99\)00149-5](https://doi.org/10.1016/S0305-0548(99)00149-5)

- 
- Xia, Y., Liu, C., Da, B., & Xie, F. (2018). A novel heterogeneous ensemble credit scoring model based on bstacking approach. *Expert Systems with Applications*, 93, 182–199. <https://doi.org/10.1016/j.eswa.2017.10.022>
- 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>
- Yu, L., Yang, Z., & Tang, L. (2016). A novel multistage deep belief network based extreme learning machine ensemble learning paradigm for credit risk assessment. *Flexible Services and Manufacturing Journal*, 28, 576–592. <https://doi.org/10.1007/s10696-015-9226-2>
- Zhang, D., Zhou, X., Leung, S. C., & Zheng, J. (2010). Vertical bagging decision trees model for credit scoring. *Expert Systems with Applications*, 37, 7838–7843. <https://doi.org/10.1016/j.eswa.2010.04.054>
- Zhang, L., Priestley, J., & Ni, X. (2018). Influence of the event rate on discrimination abilities of bankruptcy prediction models. *arXiv preprint arXiv:1803.03756*. <https://doi.org/10.5121/ijdms.2018.10101>
- Zhao, Z., Xu, S., Kang, B. H., Kabir, M. M. J., Liu, Y., & Wasinger, R. (2015). Investigation and improvement of multi-layer perceptron neural networks for credit scoring. *Expert Systems with Applications*, 42, 3508–3516. <https://doi.org/10.1016/j.eswa.2014.12.006>
- 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>