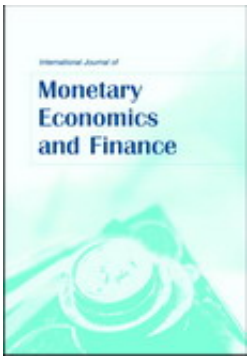


International Journal of Monetary Economics and Finance

ISSN online: 1752-0487 - ISSN print: 1752-0479
https://www.inderscience.com/ijmef

Are machine learning models more effective than logistic regressions in predicting bank credit risk? An assessment of the Brazilian financial markets

Alex Cerqueira Pinto, Alexandre Xavier Ywata de Carvalho, Mathias Schneid Tessmann, Alexandre Vasconcelos Lima

DOI: 10.1504/IJMEF.2023.10058589

Article History:

Received:	26 November 2022
Last revised:	21 February 2023
Accepted:	22 February 2023
Published online:	25 March 2024

Are machine learning models more effective than logistic regressions in predicting bank credit risk? An assessment of the Brazilian financial markets

Alex Cerqueira Pinto

Finance and Quantitative Methods,
University of Brasília,
SBS Quadra 1 Bloco A, 23, Brasília-DF, Brazil
Email: alexcerca10@gmail.com

Alexandre Xavier Ywata de Carvalho,
Mathias Schneid Tessmann*
and Alexandre Vasconcelos Lima

Brazilian Institute of Education, Development and Research (IDP),
SGAS 607 Md. 49, Brasília-DF, Brazil
Email: alexandre.carvalho@idp.edu.br
Email: mathias.tessmann@idp.edu.br
Email: alexandre.lima@idp.edu.com
*Corresponding author

Abstract: This paper seeks to investigate whether machine learning models are more efficient than logistic regressions to predict credit risk in financial institutions. Through an empirical study that develops the models and applies interpretability techniques to identify the relationships between the variables and their importance, data and economic-financial indicators from Brazilian firms in the wholesale segment are used, combined with the use of supervised machine learning. The results indicate that the model with the best predictor performance is XGBoost, with an accuracy of 0.59 and a ROC curve of 0.97 for out-of-time data. In the interpretability analysis – via sharp value – the results corroborate the importance and economic meaning of the variables. These findings confirm the improvement in the predictive capacity of the models using machine learning techniques and are useful for the financial literature and for financial market agents in general.

Keywords: credit risk measurement; machine learning to detect credit risk; credit risk in Brazilian banks; empirical evidence in finance and banking.

Reference to this paper should be made as follows: Pinto, A.C., de Carvalho, A.X.Y., Tessmann, M.S. and Lima, A.V. (2024) 'Are machine learning models more effective than logistic regressions in predicting bank credit risk? An assessment of the Brazilian financial markets', *Int. J. Monetary Economics and Finance*, Vol. 17, No. 1, pp.29–48.

Biographical notes: Alex Cerqueira Pinto is a PhD candidate in Finance and Quantitative Methods at Brasilia Universtity, Master in Economics at Brazilian Institute of Education, Development and Research – IDP and Economist at Bank of Brazil.

Alexandre Xavier Ywata de Carvalho, PhD in Statistics at Northwestern University, Researcher at Applied Economics Research Institute (IPEA) and Professor at Brazilian Institute of Education, Development and Research – IDP.

Mathias Schneid Tessmann, PhD in Economics at Catholic University of Brasilia, World Bank Consultant, Coordinator and Researcher at Brazilian Institute of Education, Development and Research – IDP.

Alexandre Vasconcelos Lima is a PhD candidate in Finance and Accounting at Fucape Business School, Master in Economics at Brazilian Institute of Education, Development and Research – IDP and Professor at Brazilian Institute of Education, Development and Research – IDP.

1 Introduction

The main contribution to the perpetuity of a Financial Institution is the evaluation of its activities from the perspective of performance and efficiency. A well-developed and efficiently functioning banking system facilitates the improvement of other business spheres in the national economy and contributes to the development of the entire country (Grmanová and Ivanová, 2018).

All financial institutions and companies of the most different branches are subject to a variety of risks during the operational cycle of their business. Knowing these risks is essential, as well as managing those to which they are most relevant in terms of their way of acting, market niche, or business scale. Damodaran (2010) describes that risk is omnipresent in almost all human activities and there is no unanimity about a definition for the term. Thus, the discussion of this topic is based on the distinction between the risk that can be objectively quantified and the subjective risk.

The credit market presents increasing competition between traditional financial institutions and the new players that are emerging, especially credit fintech. In this way, credit risk management, mainly measurement and evaluation, is fundamental for the banking segment in multiple conceptions. Credit risk is the main risk faced by financial institutions and, in general, is subject to strict supervision by national regulators, with a greater need for capital to mitigate unexpected losses as directed by the Bank of International Settlements in the so-called Basel Accords (Hull, 2012; Pesaran et al., 2006).

Meanwhile, due to their high scalability nature, machine learning methods have greater flexibility compared to more traditional economic forecasting techniques. This feature brings a better approximation of the data for measuring risk premiums. Its use for finance predicts that these types of models must be premised on stable and explanatory performance of the propositions that lead to non-compliance with the standards (Gu et al., 2018).

In this way, the use of increasingly robust models to predict default by financial institutions is a latent need, as there is growing competition in the sector due to new entrants, known as fintech. One of the main attributions of agents who work with credit risk is to develop models to predict the probability of default of an individual or company. Even with logistic regression being widely used in the risk areas of financial institutions, the process of measuring and identifying risks needs continuous improvement and development since the characteristics and risk factors of borrowing agents tend to evolve as conditions change financial and macroeconomic.

Thus, this paper seeks to test whether machine learning algorithms develop more efficient models to predict credit risk than the widely used logistic regression. Additionally, the interpretability of the variables selected for the models is analysed, that is, if the relationship between them and their influences correspond to the economic sense and their degree of importance. The algorithms Naive Bayes, Random Forest, Extreme Gradient Boosting – Xgboost, linear support vector machine – SVM and artificial neural networks (ANNs) are considered and data from publicly traded Brazilian firms in the wholesale segment are used.

The results show that some machine learning models are slightly superior to logistic regression in predicting the probability of default. The best-performing algorithm is XGBoost, with an accuracy of 0.59 and a ROC curve of 0.97. The interpretability analysis, observed through the use of the Sharpe value, shows coherence and an economic sense of the most significant variables for the model's output. These results are useful for the scientific literature that investigates the use of machine learning tools to predict financial variables by bringing empirical evidence from Brazilian banks, to bank managers and other investors who consider this information in their decisions.

In addition to this introductory section, the paper has four more sections. In Section 2, the theoretical framework on the subject is presented, Section 3 explains the data and methods used, Section 4 exposes and discusses the results and, finally, Section 5 concludes.

2 Theoretical reference

2.1 Theoretical model

To manage and measure credit risk, financial institutions use the following traditional methods as opposed to new supervised learning techniques. Among the techniques traditionally used, the Merton Model (1974) stands out in the literature. Merton (1974) presented a model for evaluating and pricing private securities through the relationship between the company's financial structure and its probability of default. In this way, the author, using the model of Black and Scholes (1973) on the option pricing theory, widely used by the financial market until today, Merton extended its application to debts and loans in general. In this way, the Merton model makes it possible to obtain the firm's default probability and, in the same context, the implicit credit spread. As described by Souza and Corrar (2010), mathematically we have:

$$E_0 = A_0.N(d_1) - D.e^{-rT}.N(d_2) \tag{1}$$

where

E_0 = Market Shareholders' Equity (PL M) at the moment 0

A_0 = Asset Market Value (A_M) at the moment 0

$N(d_1)$ = Normal distribution accumulated up to the point d_1

D = Debt face value

r = Risk-free interest rate

T = Debt duration, or CDS Term

$N(d_2)$ = Normal distribution accumulated up to the point d_2.

In this sense, the values of d_1 and d_2 are:

$$d_1 = \frac{\ln\left(\frac{A_0}{D}\right) + \left(r + 0,5\sigma_A^2\right)T}{\sigma_A\sqrt{T}} \tag{2}$$

$$d_2 = \frac{\ln\left(\frac{A_0}{D}\right) + \left(r - 0,5\sigma_A^2\right)T}{\sigma_A\sqrt{T}} \tag{3}$$

Another model widely used by financial institutions and credit bureaus is the model using the Logistic Regression technique, also known as logit analysis. This model deals with a multivariate analysis technique, appropriate for situations in which the dependent variable is categorical and assumes one of two possible outcomes, using binary marking of zero (solvent) and one (insolvent). The objective of logistic regression is to generate a mathematical function whose answer allows establishing the probability of an observation belonging to a previously determined group, due to the set of independent variables. Thus, the coefficients estimated by the regression model indicate the importance of each independent variable for the occurrence of the event (Brito et al., 2009).

Mathematically, logistic regression is described as:

$$logit\left[\theta(x)\right] = \log\left[\frac{\theta_x}{1-\theta_x}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{4}$$

where α is a constant of the model and β are the coefficients of the predictor variables.

2.2 Literature review

The scientific literature presents a multitude of studies regarding the analysis of credit risk and different methods for its evaluation and measurement. In this sense, Brito et al. (2009) built a traditional credit risk model, using the logistic regression technique, where

they used a sample of 60 publicly traded companies in the period between 1994 and 2004, classifying them as solvent or insolvent. From the bankruptcy filing registered in the Boletim Diário de Informação reports, published by Bovespa and the register of publicly-held companies of the Securities and Exchange Commission (CVM). The independent variables used were financial indices calculated from the financial statements of the companies of the penultimate year before the year of the default event. The final model showed excellent predictive power, with a ROC curve of 0.97, and was composed of the intercept and four explanatory variables, namely: i) retained earnings on assets; ii) financial indebtedness; iii) net working capital; and iv) treasury balance on sales.

Soares and Rebouças (2015) presented a study to calculate credit risk that tested several models for predicting the insolvency of publicly traded Brazilian companies, using a sample of 21 insolvent companies and 66 solvent companies, chosen according to the sectoral distribution of the first group. The following techniques were used: discriminant analysis, logistic regression, classification and regression trees (CART) and ANN, the latter with performance considerably superior to the others.

Guimarães and Moreira (2008) propose a model for predicting insolvency based on accounting indicators using discriminant analysis. The authors used a sample composed of financial and accounting information from 116 publicly traded companies from 17 different sectors, between 1994 and 2003, collected from the IBMEC Company Balance Sheet Analysis System database. The accounting indicators of the companies with default were extracted from the financial statements referring to a year before entering the state of insolvency. Thus, regarding the predictive variables of the model, the authors confirmed the discriminatory power of those that evidence financial decisions on asset structure, capital structure and cash generation.

Also in the context of credit risk, Jackson and Wood (2013) propose and test models for predicting insolvency for bankruptcy of companies based on accounting data and investigate their effectiveness. Thus, they used data from UK, with companies that failed between 2000 and 2009 considering as a population all non-financial companies listed on the London Stock Exchange (LSE), in a set of 101 companies with bankruptcy. Of the 13 models tested, the four best-performing models are contingent claims models based on European call and barrier options. Another conclusion is based on the fact that the cash flow and total debt variables present good predictive capacity.

In similar research, Luo et al. (2017) apply, from the perspective of credit risk rating, modelling with the use of Deep Learning algorithms for credit rating of global companies, which operate in different sectors, after the crisis of 2008, using the companies' Credit Default Swap (CDS) as a variable. Test results indicate that Deep Learn outperforms other algorithms and can easily rank companies' credit risk in advance.

Finally, Xia et al. (2017) proposed a credit risk scoring model based on the Extreme Gradient Boosting model, known as XGBoost. The model mainly comprises three steps that were used to calibrate the model to offer assertive classifying power. The results of the work demonstrate that the model proposed by the authors outperforms, on average, the traditional models in four performance measures: precision, error rate, area under the curve (AUC), in addition to better coverage and interpretability in the credit score.

Thus, the present work seeks to contribute to this literature by comparing the efficiency of machine learning algorithms with logistic regression for the prediction of credit risk by bringing empirical evidence to Brazilian financial institutions.

3 Methodology

3.1 Data

The creation of a model has several steps that comprise the formulation of the objectives and hypotheses of the model, data collection and treatment, creation and marking of the model's target, the pre-processing of the data, hyperparametrisation, training, testing and validation of the model.

The indicator for comparing the models to choose the best technique is done through the area under the ROC curve and the Precision of the confusion matrix. Because it is about creating and comparing models with machine learning techniques, the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology is used. This methodology brings together some of the best data mining practices so that the data processing and modelling process is as productive and efficient as possible (Tukey, 1977).

In this sense, the way machine learning algorithms learn can be classified into supervised, unsupervised and reinforcement learning. This paper used supervised learning, which provides a set of labelled data where for each observation the correct output or category is informed. In summary, supervised models present the dependent variable, also called the target, in the model training data, so the algorithm seeks to recognise a pattern of behaviour of the observations to later perform the classification. During the training and validation stages of the model, the cross-validation methodology will be used to use the best possible hyperparameters, avoid overfitting and raise the performance indicators with greater variability of the training and test data.

The data used in the work are made available by the Brazilian Securities Commission – CVM and refer to the balance sheets reported by the companies in the period from 2011 to 2019 with all the financial statements distributed by the Brazilian stock exchange – B3 and by the CVM in the DFP, FRE systems and FCA. As in Brito et al. (2009), the population used in the study, from which the sample will be selected, comprises non-financial wholesale companies registered with the CVM. The withdrawal of financial companies was due to their equity structure, distribution and characteristics of assets and liabilities being different from other non-financial companies. Thus, 8766 quarterly observations are collected from a total of 473 companies in the analysed period, organised in a stacked panel format that does not consider their temporal aspect.

The choice of companies of this nature and size occurs because they are companies with information in the public domain and easy disclosure of accounting information. The data collected in the form of accounting information from these companies were transformed into economic/financial indicators to be used as predictive variables for the models to be tested.

The list of independent variables tested in the models is described in Tables 1 and 2. These are 27 variables constructed from the companies' accounting indices and macroeconomic indices and indicators.

Table 1 Accounting ratios for use as model variables

Variable code	*Financial index*	*Index formula*
V1	General liquidity	(Current Assets + Long-Term Assets) / (Current Liabilities + Long-Term Liabilities)
V2	Current liquidity	Current Assets / Current Liabilities
V3	Dry liquidity	(Current Assets – Inventories) / Current Liabilities
V4	Immediate liquidity	Available / Current Liabilities
V5	Return on equity	Initial Net Income / Equity
V6	Return on asset	EBIT / Total Assets
V7	Return on sales	Net Income / Net Sales
V8	Asset turnover	Net Sales / Total Assets
V9	Operating margin	EBIT / Net Sales
V10	Operating profit over financial expenses	EBIT / DF
V11	Equity over assets	Shareholders' Equity / Total Assets
V12	Retained earnings on assets	(Retained Earnings + Earnings Reserve) / Total Assets
V13	Shareholders' equity over total liabilities	Equity / (Current Liabilities + Long-Term Liabilities)
V14	Total indebtedness	(Current Liabilities + Long-Term Liabilities) / Total Assets
V15	Short-term debt	Current Liabilities / Total Assets
V16	Financial indebtedness	(Current Financial Liabilities + Long-Term Financial Liabilities) / Total Assets
V17	Immobilisation of equity	Permanent Assets / Shareholders' Equity
V18	Inventories on assets	Inventories / Total Assets
V19	Net working capital	(Current Assets – Current Liabilities) / Total Assets
V20	Need for working capital	(Operating Current Assets – Operating Current Liabilities) / Total Assets
V21	Treasury balance on assets	(Current Financial Assets – Current Financial Liabilities) / Total Assets
V22	Cash balance on sales	(Current Financial Assets – Current Financial Liabilities) / Net Sales
V23	Operating cash flow over assets	Cash Flow from Operations / Total Assets
V24	Operating cash flow over total liabilities	Cash Flow from Operations / (Current Liabilities + Long-Term Liabilities)
V25	Operating cash flow on financial indebtedness	Cash Flow from Operations / (Current Financial Liabilities + Long-Term Financial Liabilities)

Source: Elaborated by authors

Table 2 List of macroeconomic variables

Code	*Financial index*	*Type*
M1	The basic interest rate over the quarter	Numeric
M2	Potential GDP Deviation – % per year	Numeric
M3	Brazil Risk (Credit default swap)	Numeric
M4	Installed Capacity Utilisation Index – UCI	Numeric
M5	General Government Primary Result	Numeric
M6	VIX – Volatility Index	Numeric
M7	Interest differential between Brazil and USA	Numeric
M8	Interest differential between Brazil and Emerging Countries	Numeric
M9	IBOVESPA profitability (Brazilian stock exchange index)	Numeric
M10	NTN-B profitability (inflation-linked bond) – Premium	Numeric
M11	GDP – Quarterly Change – %	Numeric
M12	GDP – Moving Average Change 2 years – %	Numeric
M13	GDP – Moving Average Change 3 years – %	Numeric
M14	Inflation – IPCA	Numeric
M15	Exchange Rate – Real/Dollar – Nominal	Numeric
M16	Exchange Rate – Real/Dollar – Real – Index	Numeric

Source: Elaborated by authors

The default flag in the database will be the model's dependent variable, also called the target. Thus, to avoid endogeneity in the response variable, companies identified in bankruptcy and judicial reorganisation through the registration status report of companies registered with the CVM were marked as insolvent. Even so, to indicate a predictive character to the model, the marking in the database for companies in default was carried out one year before the date of filing for bankruptcy or judicial recovery. Thus, the database was marked, in its target, as 1 (one) for default, and 0 (zero) for non-default. This target marking is considered independent and exogenous to the companies' balance sheets, ensuring no correlation between the independent variables and the response variable.

Thus, after excluding observations from the database that would prevent the creation of the model (such as absences of relevant information or missings) the final modelling database used data from 423 companies in 7734 quarterly observations, so that 384 marked were marked as default, representing 4.96% of the sample. As these are unbalanced data, models with balanced data in the proportion of 1/1 were also evaluated using the SMOTE algorithm, developed by Chawla et al. (2002). In this method, the authors developed a methodology of oversampling the minority class involving the creation of synthetic minority class examples.

3.2 Machine learning techniques

Machine learning techniques are considered one of the most recent and important advances in applied mathematics, with diverse applications in medicine, economics, finance, robotics, and various segments of industry and services. As described by

Tian et al. (2012), machine learning is a sequence of computer science, which, together with applications of statistics, gave software the ability to learn behaviour patterns, being used mainly in classification problems. and prediction.

In this way, a quick contextualisation of the techniques used will be given, such as Naive Bayes, Random Forest, Extreme Gradient Boosting – Xgboost, Linear Support Vector Machine – SVM and ANNs.

The Naive Bayes algorithm is a probabilistic binary classification algorithm widely used in machine learning. Based on Bayes' Theorem, which deals with conditional probability, that is, the probability that event A will occur, given event B. The algorithm assumes that there is independence between the variables of the model, that is, the algorithm assumes that the probabilities are conditionally independent of the target instead of calculating the value of the probabilities related to each attribute. Hence its naive name – naive (Lewis, 1998).

In this way, the Bayesian network is described as follows:

$$P\left(C=\frac{c_k}{X}=x\right)=P\left(C=c_k\right)x\frac{P\left(X=\frac{x}{C}=c_k\right)}{P_{(x)}} \tag{5}$$

$$P\left(X=x\right)=\sum_{k'=1}^{ec}P\left(X=\frac{x}{C}=c_{k'}\right)xP\left(C=c_{k'}\right) \tag{6}$$

where

$P(c|x)$ is the probability of hypothesis c given observation x. This is known as posterior probability

$P(x|c)$ is the probability of the observation given that hypothesis c is true

$P(c)$ is the probability of hypothesis h being true (regardless of the data). This is known as the prior probability of h

$P(x)$ is the probability of observation o (regardless of the hypothesis).

In the world of machine learning, one of the ways to improve the capacity of algorithms is through their combination. In this work, Random Forest and XGboost are part of this class of models. These algorithms are known as Ensemble type, that is, they combine simple models with low predictive power, to produce a single strong, robust and with greater accuracy. The main Ensemble methodologies are Bagging and Boosting.

The bagging methodology proposed used in Random Forest was proposed by Breiman (2001), and aims to reduce the variance of predictions. Several algorithms are trained separately in several resamplings with the replacement of the same training set. In general, the bagging method is based on creating multiple algorithms built for each resampled dataset and combining the predictions using means, mode, median for regression or majority vote for classification problems.

In the case of Random Forest, this model combines several decision trees and the combined values tend to be more robust than the value generated by a single model. The model builds several poorly correlated trees, where the main improvement of the combined trees is the reduction of variance. An advantage of the Random Forest technique is the ability to deal with large volumes of data and the ability to identify the

most significant variables within a set of input variables. On the other hand, as a disadvantage, the model can easily overfit the training database (overfitting), in addition to this model being difficult to interpret (James et al., 2013).

On the other hand, according to James et al. (2013), in the boosting method, the algorithms are applied sequentially so that at each iteration the applied algorithm uses the model residuals (errors) from the previous interaction as the dependent variable, instead of the response variable. Thus, XGboost is a decision tree-based boosting machine learning algorithm that uses a gradient boosting structure. This is a method that has won the most machine learning competitions on Google's Kaggle platform, often combined with deep neural networks. The most important factor behind XGBoost's success is its scalability in all scenarios due to its algorithmic optimisation. The system runs more than ten times faster than existing popular solutions on a single machine and scales to billions of examples in distributed or low-memory configurations (Chen and Guestrin, 2016).

Another technique to be tested in the paper is the Support Vector Machine – SVM. Developed by Boser et al. (1992), this is a machine learning algorithm, considered supervised learning used for classification. According to Betancourt (2005), the SVM has the following advantages:

i ease of training

ii it does not have a local optimum, as in neural networks

iii scales relatively well for data in high-dimensional spaces

iv the relationship between classifier complexity and error can be explicitly controlled

v non-traditional data such as characters can be used as input rather than feature vectors.

On the other hand, the weakness of SVM is the need for a 'good' kernel function, i.e., efficient methodologies for tuning SVM boot parameters.

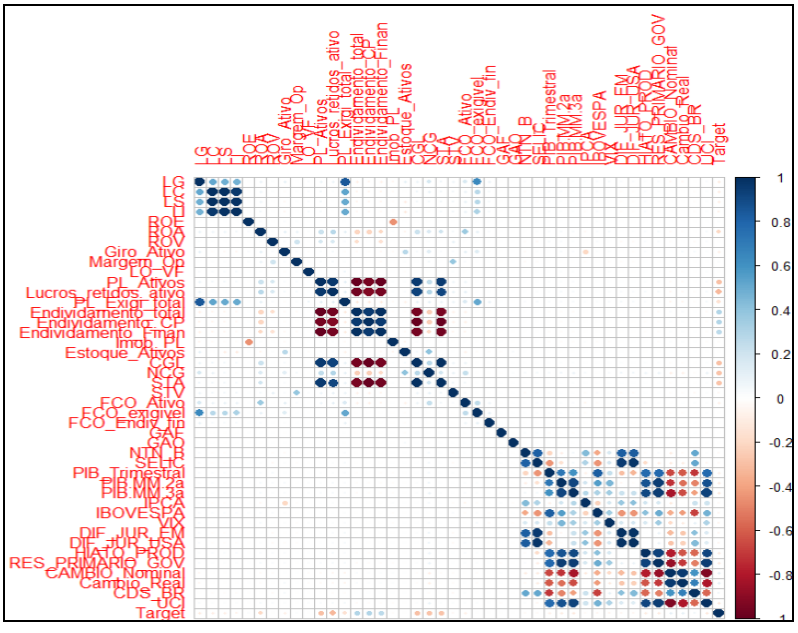
Finally, the ANN will be the last technique to be used in the project. This model was described by McCulloch and Pitts (1943), who created a system that reproduces the basic characteristics of a human neuron, the perceptron. In this way, ANNs are an information processing technique inspired by the human nervous system. As described by Haykin (2007), the human brain can be considered an extremely complex, non-linear and parallel information processing system, which performs various activities much more efficiently than computer systems.

4 Results

4.1 Choice of predictor variables

From the variables collected, at first, we evaluated the correlation relationship between the predictor variables as shown in Figure 1. This figure presents the correlation matrix of the variables initially tested in the model vs. all the others. An established cut-off point is the non-use of variables with a correlation between themselves greater than 0.6, opting for only one of the variables.

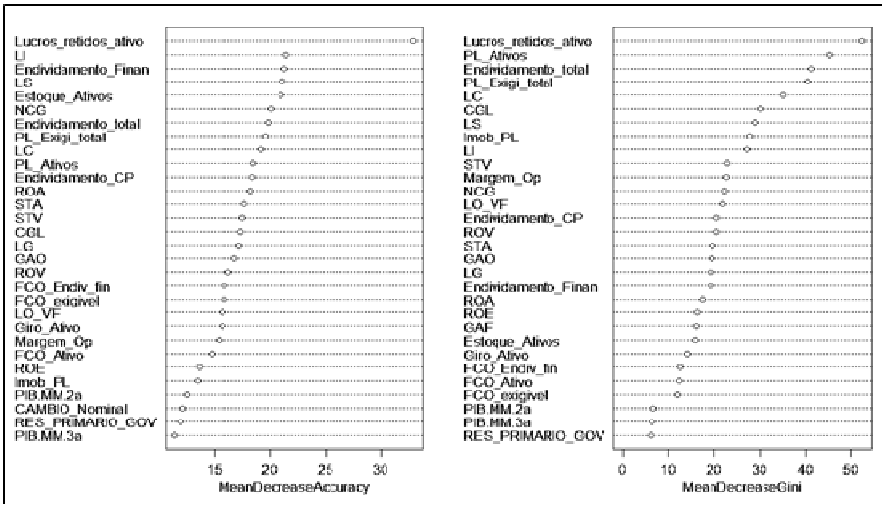
Figure 1 Pearson's correlation matrix of the predictor variables (see online version for colours)



Source: Elaborated by authors

In addition to the choice of variables, an evaluation was carried out via the relative importance of the variables, through the decision tree technique, to select those with greater predictive power to the models as highlighted in Figure 2. Another form of variable selection used was Recursive feature selection using cross-validation. Using this method, the 5 variables with the highest predictive power were: Retained Earnings/ Assets, Immediate Liquidity, Inventory/Assets, Short-Term Debt and Dry Liquidity.

Figure 2 Relative importance of variables



Source: Elaborated by authors

Through the parameters and techniques mentioned above, from the 43 initial variables, 24 variables were chosen to minimise correlations and overfitting, maximise the predictive power and maintain the model's ability to generalise. All these variables were applied and tested in all models. The final list of variables is shown in Table 3.

Table 3 List of final variables for use in models

Dry liquidity	*Asset turnover*	*Degree of financial leverage*
Treasury Balance on Sales	Need for working capital	Inflation index
Operating Profit / Financial Expenses	Return on Sales	Brazilian stock exchange index IBOVESPA
Operating Capital Flow on Liabilities to Shareholders' Equity	Operating Cash Flow on Financial Indebtedness	VIX
Shareholders' Equity on Total Liabilities	Retained Earnings on Assets	BR/Emerging Interest Differential
Inventory on Assets	Fixed Assets on Equity	Output Gap
Return on equity	Operating margin	Real exchange rate
Return on Assets	Operational Leverage Degree	CDS – Brazil

Source: Elaborated by authors

The final 24 variables presented in Table 3 will be used as predictive variables in all tested techniques. This selection was based on the analysis of the correlation of variables, the degree of the relative importance of variables and the recursive selection of variables.

4.2 Comparison of models and choice of the best performance

In the model development stage, the observation base was randomly segmented in the proportion of 80% for training the models and 20% for out-of-time testing of the models. Likewise, during the training of the models with the use of 80% of the data, the training methodology with cross-validation was used. Thus, the results presented in this session refer to the application of the models trained in the out-of-time sample.

After applying the data to the mentioned techniques, using all models with the best possible hyperparameter configuration, it was found that some algorithms using machine learning showed better predictive power than traditional logistic regression, as can be seen in Tables 4 and 5.

Because it is a very unbalanced real base, where high accuracy indicators naturally occur, precision and the ROC curve were used as criteria to choose the best model. Thus, XGBoost was chosen as the technique that presented the best performance, presenting an accuracy of 0.5895 and a ROC of 0.97 in the training data with the balanced distribution.

Table 4 was prepared with the data extracted from the application of the confusion matrix in each trained model, in the out-of-sample test data with the application of calibration of its hyperparameters.

Table 4 Algorithm performance indicators – training with level data

	Naive Bayes	*SVM – linear*	*SVM – polynomial*	*SVM – radial*	*Logistic regression*	*Neural networks*	*Random forest*	*XGBoost*
Accuracy	0.9496	0.9528	0.9509	0.9651	0.9655	0.9638	0.978	0.9838
Precision	0.6666	0.6538	0.9	0.8478	0.851	0.7258	0.9615	1
Recall	0.097	0.2098	0.1071	0.4534	0.3539	0.5357	0.6097	0.66
F1	0.17	0.3177	0.1914	0.5909	0.5	0.6164	0.7462	0.7967
ROC	0.89	0.91	0.94	0.94	0.94	0.96	0.97	0.99

Source: Elaborated by authors

Table 5 Algorithm performance indicators – training with balanced data

	Naive Bayes	*SVM – radial*	*Neural networks*	*Logistic regression*	*SVM – polynomial*	*SVM – linear*	*XGBoost*	*Random forest*
Accuracy	0.9347	0.9244	0.9166	0.9024	0.9257	0.9011	0.9586	0.9606
Precision	0.4054	0.3629	0.355	0.3251	0.3856	0.3236	0.5825	0.5798
Recall	0.5625	0.6125	0.75	0.825	0.7375	0.8375	0.90	0.8625
F1	0.4712	0.4558	0.4819	0.4664	0.5064	0.4669	0.6923	0.6934
ROC	0.87	0.88	0.90	0.91	0.91	0.92	0.97	0.97

Source: Elaborated by authors

Similarly to the previous table, Table 5 is prepared with the data extracted from the application of the confusion matrix (confusion matrix) in each trained model, with the balancing data, in the test data and out of the sample, with the application of calibration of its hyperparameters.

4.3 Application of the interpretability of the champion model as validation

For the XGBoost model, which presented the best predictive capacity, to complement this work, an analysis of the interpretability of its estimates was performed. Models with this type of technique are considered black-box because it is a non-traditional algorithm where it is not possible to observe the values of the estimated betas for the variables, as well as to perform standard statistical tests.

Machine learning models, known as black-box, are increasingly being studied to clarify the relationship between the explanatory variables and the model output. It is important to emphasise that the interpretability of the model also aims to assess the quality of the model, as it allows for the understanding of the relationship between the explanatory variables and the output. For this paper, three tools will be applied that seek to give the models interpretability:

a Sharpley value

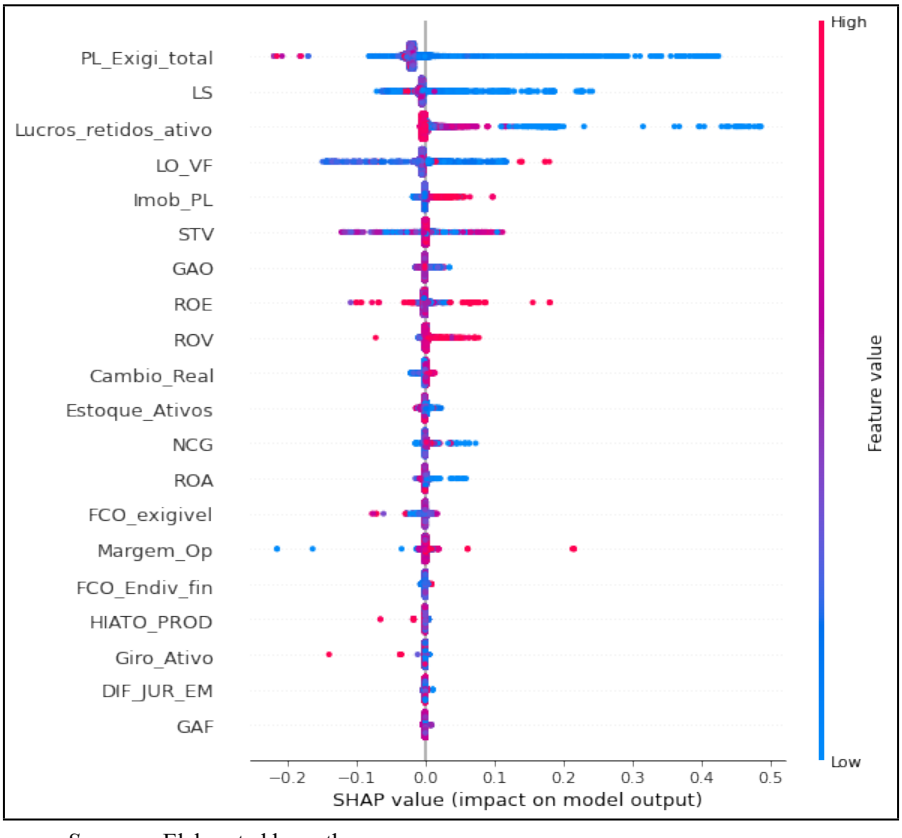
b importance

c interactions.

Based on the concept of Sharpley Value, derived from Game Theory, Lundberg and Lee (2017) developed a method applied to machine learning models that confers interpretability to the models. In this context, the Sharpe value measures the contribution of each variable in the construction of the output, that is, the fair value that each variable influences on the model's result. Thus, the relationship between these values and the values of the covariates allows for assessing the economic significance of each variable.

Figure 3 presents, in ascending order, the variables with the highest fair value, that is, the most important for the model. The observations to the right of the central line indicate that the values of each variable contribute to a greater probability of the company being in default.

Figure 3 Model interpretability analysis with Shap value (see online version for colours)

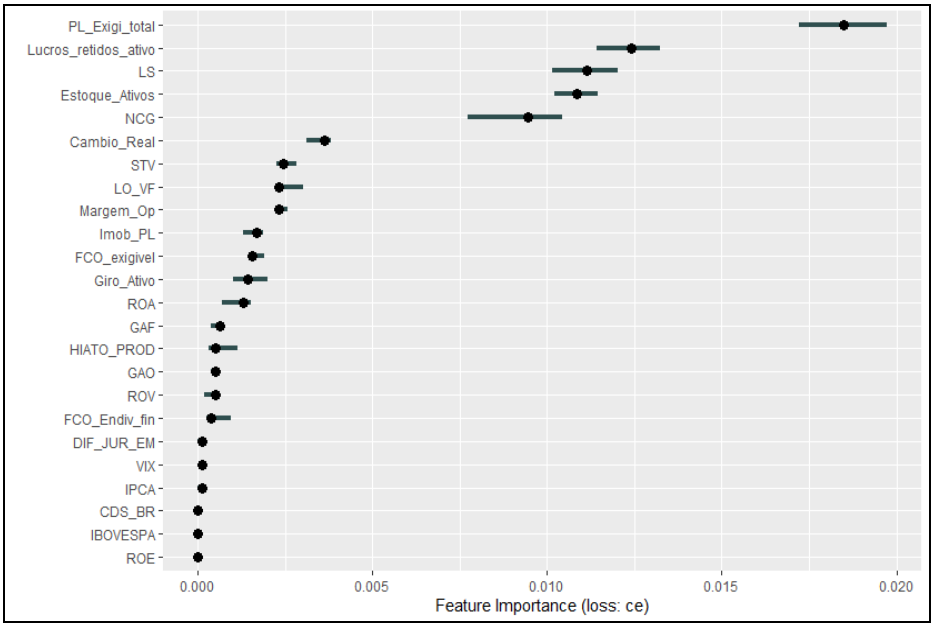


Source: Elaborated by authors

Figure 3 orders the variables by importance and corroborates the results presented by the model concerning the economic interpretation of the results. Thus, the Sharpley Value indicates an inverse relationship between the representative variables PL/Total Liabilities, Dry Liquidity and Retained Earnings/Assets and the predicted value. That is, the lower the values of these variables, the greater the probability of noncompliance.

The Sharpley Value method also allows local interpretability of the model to be performed. In this case, the results explain each predicted value individually. This method is most useful if there is a need to explain an individually predicted value. Figure 4 presents, in ascending order (from top to bottom), the most relevant variables for the model according to the classification error.

Figure 4 Importance of attributes – variables



Source: Elaborated by authors

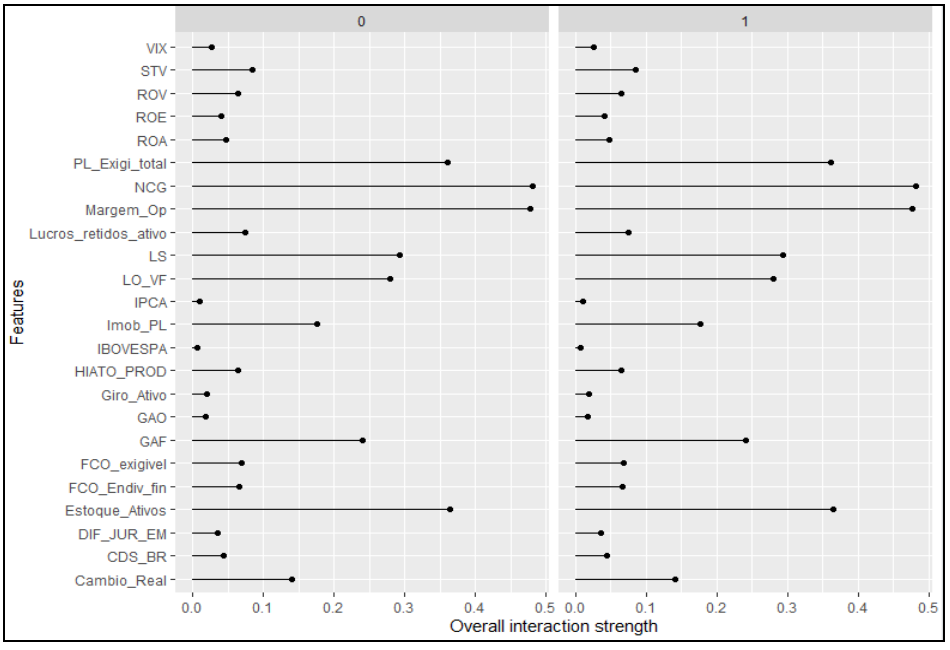
The analysis of the importance of the attributes, as shown in Figure 4, corroborates the parameters in the section for choosing the predictor variables and presents, in descending order, the most important variables for the algorithm, with emphasis on those represented by the indicators PL/Total Liabilities, Retained Earnings /Assets, Dry Liquidity, Inventory/Assets and Working Capital Need (NCG).

The output of a model is not generated only by the explanatory variables, but also through the interaction between the variables. Therefore, an evaluation was carried out through the calculation of the H-statistic of how the algorithm uses the interaction of variables to form the prediction of the model as proposed by Friedman and Popescu (2008).

The interaction measure concerns how much of the variance of $f(x)$ is explained by the interaction. The measure is between 0 (no interaction) and 1 (=100% of the variance of $f(x)$ due to interactions). For each variable, it is measured how much it interacts with any other variable. The model performs the interaction of variables as a way to increase its predictive capacity. Figure 5 denotes these interactions.

Figure 5 shows the strength of interaction between each variable used in the XGBoost model, with the other variables in the model, to formulate the predictive capacity of the algorithm. In general, it can be considered that the interaction between the variables is important in generating the model's output, especially concerning the variables PL/Total Liabilities, Working Capital Need (NCG), Retained Earnings/Assets and Inventory/ Active. In this case, it is worth detailing the interaction of these variables, as shown below. It is also worth noting that the zero indicator variables explain the predicted values based on their values, and not through interactions.

Figure 5 Variable interactions



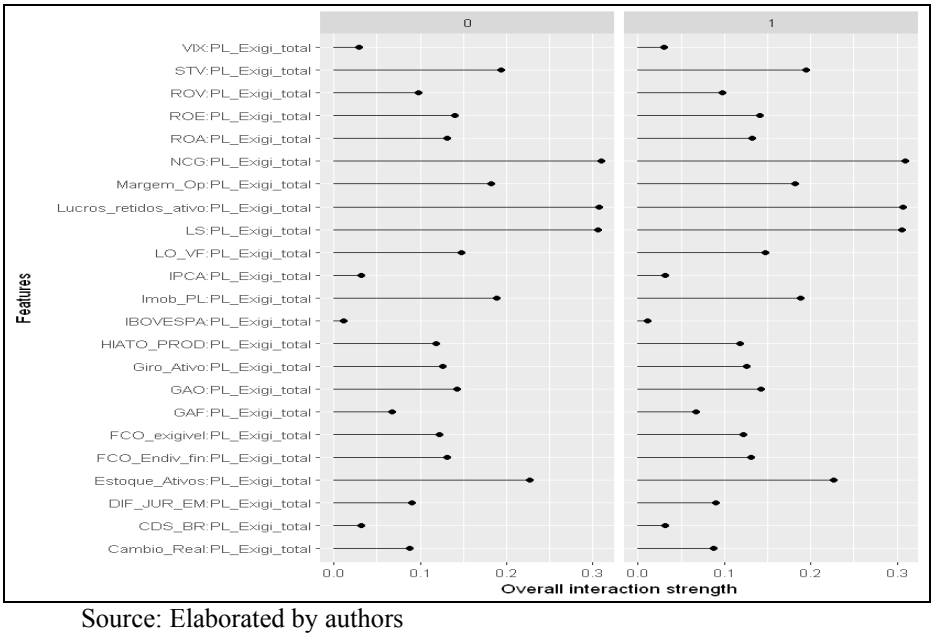
Source: Elaborated by authors

Thus, as can be seen in Figure 6, the PL/Total Liabilities variable has greater interaction with the Working Capital Need (NCG), Retained Earnings/Assets and Dry Liquidity variables. Regarding the interactions that the model performs with the Working Capital Need (NCG) variable, it has greater interaction with the Retained Earnings/Assets variable followed by the PL/Total Liability and Dry Liquidity and Inventory/Assets, as shown in Figure 7.

Figure 6 shows the strength of interaction between each variable used in the XGBoost model, with the variables PL/Total Liabilities, to compose the predictive capacity of the algorithm.

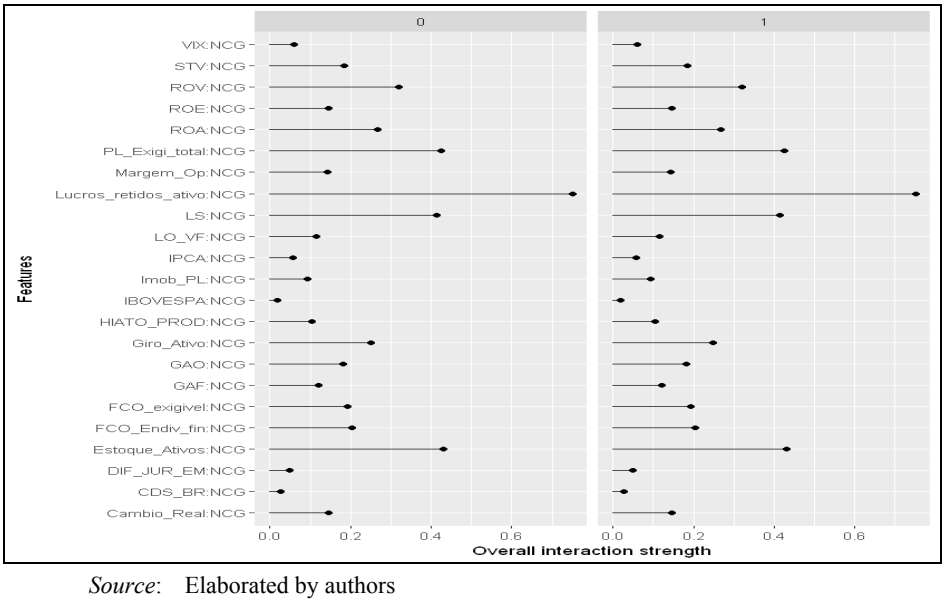
Figure 7 shows the strength of interaction between each variable used in the model with the Working Capital Need (NCG) variable for the composition of the algorithm's predictive capacity. Finally, the Operating Margin variable, unlike the other variables analysed, performs greater interaction with the STV variable, followed by the variables Retained Earnings on Assets and Ibovespa as shown in Figure 8.

Figure 6 Interaction of the PL/Total liabilities variable with the others



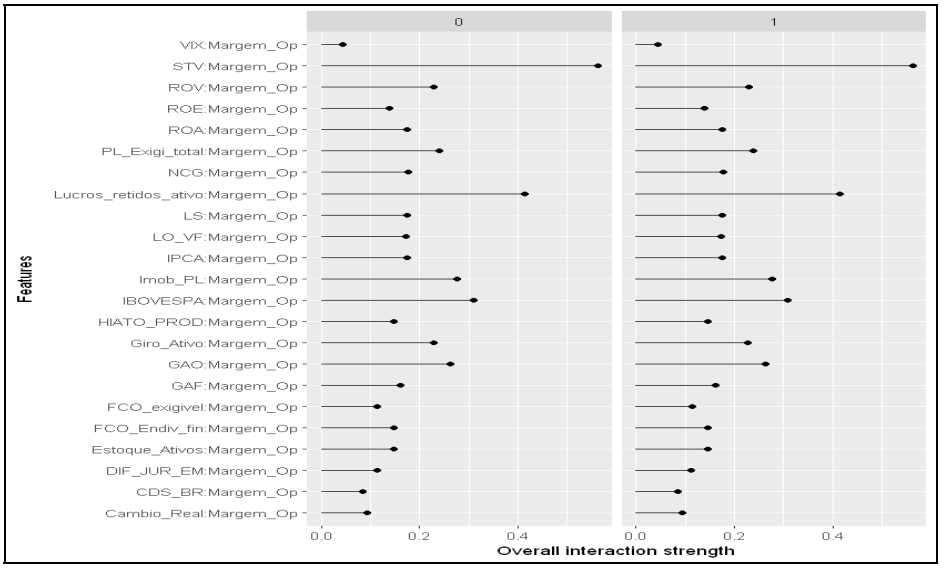
Source: Elaborated by authors

Figure 7 Interaction of the working capital need (NCG) variable with the other predictor variables



Source: Elaborated by authors

Figure 8 Interaction of the operating margin variable with the other predictor variables



Source: Elaborated by authors

The results obtained confirm the findings in the literature, in which machine learning models that use the ensemble technique, in the present study XGBoost and Random Forest, were those that presented the best results, considering the four evaluation metrics. The improvement obtained allows for greater efficiency of the financial institution, which needs to assess credit risk and discriminate bad from good payers. In this way, it is possible to improve the decision-making of managers and optimise the provision of financial services.

Despite being black-box techniques, it is possible to evaluate the features and relevance of each one to the model, in addition to the interaction of one variable with another, to verify the interactions of the most important features for the model and the other independent variables.

5 Conclusion

As an effect of the growing competition between traditional banks and fintech, it is increasingly important for credit providers to use robust models, capable of improving the ability to estimate and assess credit risk for different customer profiles, both individuals and corporations. The use of machine learning techniques, due to its large computational scale and pattern recognition capacity, is increasingly important for a better classification of the clients that the financial institution wants to work with, given its risk-taking capacity.

This paper sought to compare different types of models for predicting credit risk, measured by the probability of default, to verify the performance of the models using machine learning techniques in terms of their predictive capacity, comparing them to the logistic regression model. For this purpose, a database was used with macroeconomic information and economic and financial balance sheets of firms in the wholesale

segment, made available by the Brazilian Securities and Exchange Commission, between 2011 and 2019, with a total of 7734 observations.

Thus, it was verified that, from the data presented, the predictive capacity of some machine learning models presents a significantly superior predictive performance than the traditional logistic regression. Thus, the XGBoost model was chosen as the one that presented the best performance – ROC of 0.97 for balanced data – among the models tested.

Regarding the most important variables for the model, the most relevant ones are in line with what is stated in the literature when pointing out the choice of the company in its formation of debt, the relationship between equity and total liabilities, the ability to retain profits vs. its total assets and also the management of the company's liquidity.

For the model's interpretability assessments, they showed consistency between the direction of the values of the variables and their economic sense, as well as pointing to the relationship of the most important variables and their main interactions, such as PL / Total Liability, Retained Earnings / Assets, Dry Liquidity, Inventory / Assets and Working Capital Need (NCG). It is also verified that the nominal value of the variable, its magnitude, presents a coherent economic sense for the attribution of the fair value (Sharpe value) towards the default or non-default classification, corroborating, once again, the economic sense of the model.

It should be noted that, although the Sharpe value and more interpretability indicators are still embryonic and little used in finance, banks and supervisors for model validation, it is evident that they have expanded their scope to interpret black-box algorithms where they are not visible. The weights – betas – of each variable in conjunction with the final prediction.

The findings of this paper should not be underestimated, they are useful for the scientific literature that investigates credit risks, for the literature that studies the efficiency in the use of machine learning algorithms by banks, as well as for the financial market agents that seek to make more efficient decisions. As a suggestion for future research, we can highlight the measurement of the efficiency of deep learning algorithms as predictors of credit risk.

References

Banco Central Do Brasil – Bacen (2009) *Resolução, C.M.N. Nº 3.721,* Dispõe sobre a implementação de estrutura de gerenciamento do risco de crédito.

Betancourt, G.A. (2005) 'Las máquinas de soporte vectorial (SVMs), Scientia et Technica, 1(27).

Black, F. and Scholes, M. (1973) 'The pricing of options and corporate liabilities', *Journal of Political Economy*, Vol. 81, No. 3, pp.637–654.

Boser, B.E., Guyon, I.M. and Vapnik, V.N. (1992) 'A training algorithm for optimal margin classifiers', *Proceedings of the Fifth Annual Workshop on Computational Learning Theory,* July, Pittsburgh, Pennsylvania, USA, pp.144–152.

Breiman, L. (2001) 'Random forests', *Machine Learning*, Vol. 45, No. 1, pp.5–32.

Brito, G.A.S., Assaf Neto, A. and Corrar, L.J. (2009) 'Sistema de classificação de risco de crédito: uma aplicação a companhias abertas no brasil', *Revista Contabilidade and Finanças*, Vol. 20, No. 51, pp.28–43.

Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) 'SMOTE: synthetic minority over-sampling technique', *Journal of Artificial Intelligence Research*, Vol. 16, pp.321–357.

Chen, T. and Guestrin, C. (2016) 'Xgboost: A scalable tree boosting system', *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, August, San Francisco, California, USA, pp.785–794.

Damodaran, A. (2010) *Avaliação de investimentos: ferramentas e técnicas para a determinação do valor de qualquer ativo*, Qualitymark.

Friedman, J.H. and Popescu, B.E. (2008) 'Predictive learning via rule ensembles', *The Annals of Applied Statistics*, Vol. 2, No. 3, pp.916–954.

Grmanová, E. and Ivanová, E. (2018) 'Efficiency of banks in Slovakia: measuring by DEA models', *Journal of International Studies*, Vol. 11, No. 1, pp.257–272.

Gu, S., Kelly, B. and Xiu, D. (2018) *Empirical Asset Pricing Via Machine Learning* (No. w25398) National Bureau of Economic Research.

Guimarães, A. and Moreira, T.B.S. (2008) 'Previsão de insolvência: um modelo baseado em índices contábeis com utilização da análise discriminante', *Revista de Economia Contemporânea*, Vol. 12, No. 1, pp.151–178.

Haykin, S. (2007) *Redes Neurais: Princípios e Prática*, Bookman Editora.

Hull, J. (2012) *Risk Management and Financial Institutions, Web Site* (Vol. 733), John Wiley & Sons, Toronto, Canada.

Jackson, R.H. and Wood, A. (2013) 'The performance of insolvency prediction and credit risk models in the UK: a comparative study', *The British Accounting Review*, Vol. 45, No. 3, pp.183–202.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013) *An Introduction to Statistical Learning*, Springer, New York, Vol. 112, p.18.

Lewis, D.D. (1998) 'Naive (Bayes) at forty: the independence assumption in information retrieval', *European Conference on Machine Learning*, April, Springer, Berlin, Heidelberg, pp.4–15.

Lundberg, S.M. and Lee, S.I. (2017) 'A unified approach to interpreting model predictions', *Advances in Neural Information Processing Systems*, Vol. 30, pp.4765–4774.

Luo, C., Wu, D. and Wu, D. (2017) 'A deep learning approach for credit scoring using credit default swaps', *Engineering Applications of Artificial Intelligence*, Vol. 65, pp.465–470.

McCulloch, W.S. and Pitts, W. (1943) 'A logical calculus of the ideas immanent in nervous activity', *The Bulletin of Mathematical Biophysics*, Vol. 5, No. 4, pp.115–133.

Merton, R.C. (1974) 'On the pricing of corporate debt: the risk structure of interest rates', *The Journal of Finance*, Vol. 29, No. 2, pp.449–470.

Pesaran, M.H., Schuermann, T., Treutler, B.J. and Weiner, S.M. (2006) 'Macroeconomic dynamics and credit risk: a global perspective', *Journal of Money, Credit and Banking*, Vol. 38, No. 5, pp.1211–1261.

Soares, R.A. and Rebouças, S.M.D.P. (2015) 'Avaliação do desempenho de técnicas de classificação aplicadas à previsão de insolvência de empresas de capital aberto brasileiras', *Revista ADM. MADE*, Vol. 18, No. 3, pp.40–61.

Souza, Ê.B.M. and Corrar, L.J. (2010) 'O uso do modelo de merton para obtenção de spreads de crédito: uma proposta de implementação simplificada', *Sociedade, Contabilidade e Gestão*, Vol. 5, No. 1, pp.1–20.

Tian, Y., Shi, Y. and Liu, X. (2012) 'Recent advances on support vector machines research', *Technological and Economic Development of Economy*, Vol. 18, No. 1, pp.5–33.

Tukey, J.W. (1977) *Exploratory Data Analysis*, Vol. 2, pp.131–160.

Xia, Y., Liu, C., Li, Y. and Liu, N. (2017) 'A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring', *Expert Systems with Applications*, Vol. 78, pp.225–241.