ISSN 1808-057X
DOI: 10.1590/1808-057x202010360

ORIGINAL ARTICLE

Evaluating company bankruptcies using causal forests

Wanderson Rocha Bittencourt[1]
https://orcid.org/0000-0003-3417-2225
Email: wandersonrochab@yahoo.com.br

Pedro H. M. Albuquerque[1]
https://orcid.org/0000-0002-1415-716X
Email: pedroa@unb.br

[1] Universidade de Brasília, Faculdade de Administração, Contabilidade, Economia e Gestão de Políticas Públicas, Departamento de Administração, Brasília, DF, Brazil

Received on 08.06.2019 – Desk acceptance on 09.09.2019 – 3rd version approved on 01.29.2020 – Ahead of print on 07.10.2020
Associate Editor: Fernanda Finotti Cordeiro Perobelli

ABSTRACT

This study sought to analyze the variables that can influence company bankruptcy. For several years, the main studies on bankruptcy reported on the conventional methodologies with the aim of predicting it. In their analyses, the use of accounting variables was massively predominant. However, when applying them, the accounting variables were considered as homogenous; that is, for the traditional models, it was assumed that in all companies the behavior of the indicators was similar, and the heterogeneity among them was ignored. The relevance of the financial crisis that occurred at the end of 2007 is also observed; it caused a major global financial collapse, which had different effects on a wide variety of sectors and companies. Within this context, research that aims to identify problems such as the heterogeneity among companies and analyze the diversities among them are gaining relevance, given that the sector-related characteristics of capital structure and size, among others, vary depending on the company. Based on this, new approaches applied to bankruptcy prediction modeling should consider the heterogeneity among companies, aiming to improve the models used even more. A causal tree and forest were used together with quarterly accounting and sector-related data on 1,247 companies, 66 of which were bankrupt, 44 going bankrupt after 2008 and 22 before. The results showed that there is unobserved heterogeneity when the company bankruptcy processes are analyzed, raising questions about the traditional models such as discriminant analysis and logit, among others. Consequently, with the large volume in terms of dimensions, it was observed that there may be a functional form capable of explaining company bankruptcy, but this is not linear. It is also highlighted that there are sectors that are more prone to financial crises, aggravating the bankruptcy process.

Keywords: causal forest, causal tree, heterogeneity, financial crisis, bankruptcy.

Correspondence address

Wanderson Rocha Bittencourt
Universidade de Brasília, Faculdade de Administração, Contabilidade, Economia e Gestão de Políticas Públicas, Departamento de Ciências Contábeis e Atuariais
Campus Universitário Darcy Ribeiro, Bloco A-2 – CEP 70910-900
Asa Norte – Brasília – DF – Brazil

1. INTRODUCTION

Empirical studies often focus on the structure, causality, or treatment of a phenomenon of interest. In economics, for example, some studies seek to analyze the effects of an economic policy on economic development and employment, among others. However, there are unobservable conditions that make the strategy unviable, with it obtaining undesirable effects (Belloni, Chernozhukov, & Hansen, 2014a).

Within this setting, computational resources are gaining space, and their application in contexts such as economics and finance is inevitable. Computer systems are helping in the analysis of large databases (big data) in which the conventional statistical tools, such as regression analysis, present results that fall short of those of other tools (Varian, 2014, 2016).

With the traditional statistical tools (regressions), data manipulation and subsequent predictive potential are restricted, particularly to linear models, and they do not capture the relationships with other behaviors. Along this same line of thinking, the empirical studies generally report their estimates based on a single model, leaving part of the results unexplained by the functional specification that would normally lead to different punctual results (Athey & Imbens, 2015).

One solution for such estimation problems would be machine learning (ML) tools, for example techniques such as decision trees, support vector machines (SVMs), artificial neural networks (ANNs), and deep learning, among others, which present better results for more complex models, concentrating on high computational performance, as well as dealing with the presence of restrictions regarding linear or non-linear functional relationships (Varian, 2014).

With this range of possibilities, research has been developed using ML techniques for portfolio selection (Montenegro & Albuquerque, 2017), analyzing exchange rate predictions with SVM (Yaohao & Albuquerque, 2019), cryptocurrency performance prediction (Yaohao, Albuquerque, Camboim de Sá, Padula, & Montenegro, 2018), stock and option pricing models (DeSpiegeleer, Madan, Reyners, & Schoutens, 2018), building non-parametric non-linear prediction models for credit risk (Khandani, Kim, & Lo, 2010), and for financial manager selection, given that this tool serves as support when deciding the best choice of future fund administrators (Ludwig & Piovoso, 2005).

Supervised learning techniques (ML) thus focus on guiding the models based on a dataset (Athey, 2015). They also extrapolate, presenting more reliable results when the data are heterogeneous and the functional form cannot be observed. Thus, the various ML methods are more effective for problems related to prediction (Athey & Imbens, 2016), in this case of company bankruptcy.

The possibility of non-linear relationships between the variables constantly used in bankruptcy prediction may present greater accuracy with ML techniques (Tsai, Hsu, & Yen, 2014). These variables are treated as homogeneous and sometimes they are not, causing interpretations risks, primarily of the causal and imprecise effects. Debt ratios, for example, present distinct characteristics when their components are analyzed individually, explaining the heterogeneity among companies, and bringing a new perspective to studies that use such variables (Boot & Thakor, 1997; DeMarzo & Fishman, 2007; Park, 2000). Based on this, it is suspected that these characteristics may be extended to the other indicators used in bankruptcy analysis.

The use of non-parametric approaches, such as the causal forest (CF), would facilitate an understanding of the heterogeneity, enabling a flexible model with high levels of interactions and dimensions (Athey & Imbens, 2016; Wager & Athey, 2018). This approach thus enables the construction of valid confidence intervals to analyze the treatment, even considering a high number of variables in relation to the sample size.

CFs are gaining more prominence, since techniques such as K-nearest neighbor (KNN) would present limitations regarding the number of variables, raising the number of dimensions (Zhang & Zhou, 2007); that is, a greater quantity of variables would cause imprecision regarding the distance metric used, generating inaccurate estimates. Another option would be long short-term memory (LSTM); however, this methodology would be more indicated in cases of long time series, since it is based on the principle of the temporal evolution of the variables for classification (Hochreiter & Schimidhuber, 1997), and does not provide relevant results in this research, since the longest series would be five years.

In general terms, maximizing the predictability of company bankruptcy, especially after periods of deterioration, such as in a financial crisis, is gaining greater relevance. In such periods, a government intervention, for example, helping companies that are more prone to bankruptcy, avoiding decreases in employment and income for the region, would be more beneficial, reducing the regional effects of the recession.

The CF proposed by Athey and Imbens (2016) and Wager and Athey (2018) would thus resolve this problem, facilitating the analyses. In this methodology, the tree looks for groups in which the average effects of the treatment differ most. The search would be for an individualized treatment, balancing both conditions. First, the tree seeks to find where the effects of the treatment differ most and then it estimates the effects of the treatment more accurately. Moreover, using computational methods, the honesty condition is inserted, in which the sample is subdivided to train the tree (training sample), followed by the application (validation sample). Finally, each one of the leaves is estimated, analyzing the difference between the means of the treatment and control, that is, the mean from observing a company with bankruptcy characteristics.

It is within this context that this study seeks to explore the CF methodology, aiming to identify a set of relevant variables relating to company bankruptcy and find behavioral patterns in the data on companies that presented bankruptcy. The most common models, discriminant analysis and logit, are the most widely used and, when treating bankruptcies, the use of CFs is still at an early stage, with few applications, and this research thus helps future studies on company bankruptcy.

2. THEORETICAL FRAMEWORK

The studies on bankruptcy generate numerous relevant results, especially regarding capital structure, indicators used, and market sensitivity. In regard to capital structure, debt concentration enables fewer transaction costs involving the renegotiation of values. When presenting a recovery plan to a lower volume of creditors, these are more likely to accept, as they run risks of greater losses if liquidation occurs. There is also the possibility of a change in ownership, resulting in a reduction in credibility and increasing the probability of liquidation (Ivashina, Iverson & Smith, 2016). Also in the context of leverage, riskier structures are more prone to resorting to a bankruptcy process. This probability is reduced when there is a considerable amount of debts with real guarantees (Jostarndt & Sautner, 2010).

Presenting real solid guarantees to creditors, such as fixed assets, can help reduce the bankruptcy process, since these guarantees would be enough to honor the debts. However, keeping a high volume of this type of asset would compromise the company's liquidity. There is thus a negative relationship between the firm's liquidity and bankruptcy risk, a relationship that does not appear to be linear (Brogaard, Li, & Xia, 2017). For the Italian context, in which the reorganization and liquidation process mirrors chapters 7 and 11 of Title 11 of the regulations on bankruptcy and bankrupt companies in the United States Code (https://uscode.house.gov/browse/prelim@title11&edition=prelim), a company, when it falls into the reorganization process, produces an increase in interest on bank financing, directly reflected in its investments (Rodano, Serrano-Velarde, & Tarantino, 2016). Regarding the profitability indicators, such as return on equity (ROE) and return on assets (ROA), a rise in the latter of more than 15% can indicate a greater propensity for failure, being driven by the cash flow risk combined with internal and costly financing. Other results have shown that low leverage represents a higher probability of bankruptcy, possibly reflecting the low volume of credit (Giordani, Jacobson, Schedvin, & Villani, 2014).

The market is sensitive to company bankruptcy. A bankruptcy announcement informs the market of the accounting structure of the firm with difficulties, as well as its cash flows, generating two possible effects: contagion and competition (Benmelech & Bergman, 2011; Helwege & Zhang, 2016; Hertzel, Li, Officer, & Rodgers, 2008; Hertzel & Officer, 2012; Jorion & Zhang, 2007; Lang & Stulz, 1992).

The market thus understands that similar companies may be experiencing the same problems, this effect being known as contagion. On the other hand, a bankruptcy announcement conveys information about how good the remaining companies are, generating an expectation of wealth redistribution in the segment, this effect being known as competitive (Lang & Stulz, 1992). There is also the possibility of collateral effects, reducing the value of similar assets in the secondary market, generating a disequilibrium in supply and demand (Benmelech & Bergman, 2011).

There is also the expectation of market sensitivity, where the average price of stocks of companies in the same segment presents a negative reaction, that is, a drop, which may be a reflection of the contagion effect (Lang & Stulz, 1992).

2.1 Bankruptcy and ML

Given the importance of the bankruptcy issue, the studies that aim to predict it have grown, especially in recent years. Comparisons between ML methodologies (SVM, ANN, weighted least squares [WLS], and decision tree, among others) and the traditional methodologies (discriminant analysis and logit) are inevitable, with the results indicating the superiority of the computational techniques.

Min and Lee (2005) used SVM for predicting bankruptcy and a promising response was identified when comparing it with the most widespread methodologies in the literature, such as discriminant analysis and logit, revealing SVM to be superior in terms of predictive capacity, once the parameters were estimated.

Regarding the selection of financial indicators for bankruptcy prediction, Yang, You, and Ji (2011) used PLS and found it was better at predicting compared to the other traditional techniques, as well as observing a complex and non-linear relationship in the parameters.

Tsai et al. (2014) compared various ML methodologies, such as the decision tree, ANN, and SVM, and found that the ML models are better at predicting than the traditional metrics. Among these, SVM presented the best results compared with the other models studied, presenting intermediate performance. Comparing the Gaussian model with SVM and the logit model, better predictions were found with the Gaussian process than with SVM and logit, as well as slightly higher accuracy of SVM compared to logit (Antunes, Ribeiro, & Pereira, 2017).

Barboza, Kimura, and Altman (2017) compared various methodologies with ML and concluded that these present a substantial improvement in bankruptcy prediction, with around 10% more precision, especially when they include, besides the variables proposed by the Altman z score, some complementary financial indicators.

In general, when the traditional methodologies are compared with the ML ones, the latter are shown to be superior. However, when analyzing the results among the ML techniques, the conclusions are still contradictory, depending on the variables used.

3. METHODOLOGY

Various models have been employed in finance with the aim of identifying the next companies to fail. Within the context of conventional analyses, the models used, discriminant analysis (Altman, 1968) and logit (Ohlson, 1980), among others, primarily depend on a functional form pre-established by the researcher that is limited to the scope of the methodology. In machine learning, however, there may be the extrapolation imposed by the models, achieving more satisfactory results.

This requires the input or independent variables – $x \in R$ (profitability, liquidity, leverage, and gross domestic product [GDP], among others) – and the result or dependent variable – $y \in R$ or $y \in [0; 1]$, bankrupt or not bankrupt – with the aim of learning how the inputs explain company bankruptcy. The results may be non-linear models (relationship suggested in the studies of Giordani et al. [2014] and Brogaard et al. [2017]).

Other methodologies have been tested over the years; however, in many cases, the focus has only been on the use of the methodologies, and not on a robust analysis of the results found. A summary of these models can be observed in Table 1.

Table 1
Some models used in bankruptcy prediction

Generic model	Specific model	Some authors who have used it
Discriminant analysis	Basic	FitzPatrick (1932)
	Multivariate	Altman (1968), Lennox (1999), Min and Lee (2005), Cho, Kim, and Bae (2009), Lee and Choi (2013), Barboza et al. (2017), García, Marqués, Sánchez, and Ochoa-Domínguez (2017)
Logit	Basic	Ohlson (1980), Lennox (1999), Min and Lee (2005), Cho et al. (2009), Premachandra, Bhabra, and Sueyoshi (2009), Tseng and Hu (2010), Antunes et al. (2017), Barboza et al. (2017), García et al. (2017)
	Squared interval logit	Tseng and Hu (2010)
Probit	Basic	Zmijewski (1984), Lennox (1999)

Table 1
Cont.

Generic model	Specific model	Some authors who have used it
Neural networks	Basic	Pendharkar (2005), Chauhan, Ravi, and Chandra (2009), Cho et al. (2009), Tseng and Hu (2010), Tsai et al. (2014), Barboza et al. (2017)
	Reverse propagation	Lee and Choi (2013)
	Multilayer	Zmijewski (1984), Lennox (1999)
	Radial base function network	Tseng and Hu (2010)
	Evolution trained wavelet	Chauhan et al. (2009)
	Interactive model with weight*	Cho et al. (2009)
	Threshold variation	Pendharkar (2005)
Decision tree	Basic	Min and Lee (2005), Cho et al. (2009), Tsai et al. (2014)
Support vector machine	Basic	Min and Lee (2005), Yang et al. (2011), Tsai et al. (2014), Antunes et al. (2017), García et al. (2017)
	Linear	Barboza et al. (2017)
	Radial	Barboza et al. (2017)
Data envelopment analysis	Basic	Cielen, Peeters, and Vanhoof (2004), Premachandra et al. (2009), Premachandra, Chen and Watson (2011)
Gaussian process	Basic	Antunes et al. (2017)

Note: *The nomenclatures used by the authors were kept.*
**Cho et al. (2009) created the interactive model with weights based on the application of various bankruptcy prediction methodologies.*
Source: *Elaborated by the authors.*

However, problems are directly encountered regarding (i) the high volume of dimensions and (ii) heterogeneity. Non-parametric approaches that seek to analyze heterogeneous effects perform well in applications with small quantities of variables (Wager & Athey, 2018). In the ML literature, there is a variety of effective methods, the most popular of which – regression tree, random forest, and SVM, among others – imply modeling relationships between the attributes and the results (Athey & Imbens, 2016).

Among the possibilities for analyzing the effect of the 2007 financial crisis, one solution would be to include an interaction dummy; however, the models became even more complex, resulting, in this research, in more than 80 variables. These variables could be chosen using the least absolute shrinkage and selection operator (Lasso) and the post-Lasso, as will be seen below. However, we would encounter linear models, since they would be estimated by ordinary least squares (OLS). Using SVM would also be an option, but it would be limited to the non-exploration of the unobserved characteristics (particularities) of the companies. The tree and CF proposal are more recommended in this context, since they would enable the conditions to observe the most latent bankruptcy characteristics, considering the particularities of each set of companies.

3.1 Conditional Treatment

In the literature on machine learning based on prediction, the regression tree presents characteristics that are little different from the other methods, producing partitions of the population based on the variables so that all the units of a partition receive the same prediction (Athey & Imbens, 2016).

The proposal of this study would thus be to apply an incipient methodology in the context of finance, especially regarding bankruptcy evaluation, analyzing its characteristics. Thus, the studies of Athey and Imbens (2016) and Wager and Athey (2018) were applied to CFs.

CFs have properties that provide impartiality and asymptotic normality, producing a partition of the population according to the variables in which all the partitions received the same prediction. Formalizing the problem based on Athey and Imbens (2016), we have N units with $i = 1..., N$, with there being a pair for each unit $Y_i(0);Y_i(1)$, and a causal effect given by $t_i = Y_i(1)$ $Y_i(0)$. We also denote a binary indicator $W_i \in \{0,1\}$ with Wi = 0, indicating that it did not receive the treatment, and Wi = 1, which did receive it; we thus have:

$$Y_i^{obs} = Y_i(W_i) = \begin{cases} Y_i(0) \; if \; W_i = 0 \\ Y_i(1) \; if \; W_i = 1 \end{cases} \quad \boxed{1}$$

We also have X_i as a vector composed of K variables not affected by this treatment, thus generating a set of observations composed of Y_i^{obs}, W_i, X_i with $i = 1, \ldots, N$, this being an independent and identically distributed sample. It is also assumed that the observations can be exchanged and, in a randomized experiment with constant treatment attribution probabilities, $e(x) = p$ for the values of x, where the probability of the marginal effect of the treatment is given by $p = pr(W_i = 1)$ and that of the conditional treatment is given by $e(x) = pr(W_i = 1 | X_i = x)$. We thus arrive at:

$$W_i \perp (Y_i(0), Y_i(1)) \mid X_i \tag{2}$$

The conditional average treatment effect (CATE) is therefore:

$$\tau(x) \equiv E\,[Y_i(1) - Y_i(0) | X_i = x\,] \tag{3}$$

With this, Athey and Imbens (2016) obtained more precise estimates for the conditional average treatment effect, that is, $\hat{\tau}(.)$, in which $\tau(x)$ is based on the partitioning of resources, not varying in the partitions. The treatment is randomly attributed in the associated subpopulations by $X_i = x$, indicating that, once all the observable characteristics of individual i are known, the status of the treatment does not generate extra information about its possible results.

3.2 Post-Lasso

One simple possibility for analyzing the conditional effect related to some treatment and the interactions of its effect can be carried out using the Lasso (procedure adopted to choose the relevant variables in a regression model). We thus have the following model:

$$Y_i = \alpha + \beta_w W_i + \beta_x X_i + \beta_{xw} X_i W_i + \epsilon_i \tag{4}$$

So, if CATE is the true model, it can be written as follows:

$$\tau(x) = E|[Y_i | X = x, W = 1] - E[Y_i | X = x, W = 0] = \beta_w + \beta_{xw} x \tag{5}$$

Equation 5 implies different subpopulations indexed by $X_i = x$, having different effects for $\beta_{xw} \neq 0$. This approach is very common when the dimensions of the variables are small ($p = dim(X_i)$), using OLS. However, the problem increases as p grows and tends toward $p > n$, making the application of OLS unviable. The acceptable solution would thus be to apply the Lasso and subsequently the post-Lasso, choosing the variables that best explain the dependent variable using OLS. These procedures present advantageous properties when the regularization parameters are chosen appropriately (Belloni, Chernozhukov, & Hansen, 2014b; Belloni et al., 2014a), as well as presenting impartiality and asymptotic normality.

3.3 CF

With the possibility of a large size, one solution would be the CF. In a broad context, regression trees and forests can be considered neighbors, using an adaptive metric in the approximations. Generally, these types of methods use the Euclidian distance to analyze the closest neighbors. Decision trees can present narrower leaves throughout the directions in which the sign changes quickly, and longer ones in other directions. Thus, a causal tree can be built that resembles the regression tree, finding a point at which the high dimensionality does not cause as much of a problem for the estimates (Wager & Athey, 2018).

For this construction, suppose that there are independent samples (X_i, Y_i) of a regression tree. The space is then divided until partitioning it into a set of leaves L containing only training samples. Given a point x, the prediction value $\hat{\mu}(x)$ is evaluated, identifying leaf $L(x)$, which contains x, establishing:

$$\hat{\mu}(x) = \frac{1}{|\{i : X_i \in L(x)\}|} \sum_{\{i : X_i \in L(x)\}} Y_i \tag{6}$$

CFs are adaptive and flexible, making them efficient for estimating local parameters, such as the application of the CATE (Athey, Tibshirani, & Wager, 2019). Locally weighted estimators are calculated; that is, the effects of the treatment on a specific target $X_i = x$ are estimated, giving greater weights to the most relevant observations. The main benefit would be the greater efficiency in choosing the most important dimensions, reducing the dimensionality problem. By incorporating the conditional treatment (CATE), we have:

$$\hat{\mu}(x) = \frac{1}{|\{i : W_i = 1 \in L(x)\}|} \sum_{\{i : W_i = 1\, \in L(x)\}} Y_i - \frac{1}{|\{i : W_i = 0 \in L(x)\}|} \sum_{\{i : W_i = 0\, \in L(x)\}} Y_i \tag{7}$$

The CF thus generates a set B of causal trees, in which each one produces an estimate $\hat{\tau}(x)$. The forests thus aggregate their predictions calculating the mean $B^{-1}\sum_{b=1}^{B}\hat{\tau}_b(x)$. Using the output mean of many trees, the mean effect of the conditional can also be calculated. These procedures ignore the information about the result, since they set sample divisions, calling them honesty, producing large leaves with asymptotic normality in each one. It warrants mentioning that no item of data was wasted, thus satisfying the honesty properties.

The sample divisions, also known as sample partitioning, are made, generating an estimation sample and a test sample. After this procedure, the results are estimated and a cross-validation process is carried out in which it is possible to predict the punctual estimates of the effect of the treatment on the estimative sample. Also in this procedure, the tree is trimmed based on its level of complexity (complexity parameter).

With this, it is assumed that the individual causal trees in the forest are random subsamples of treatment examples (Athey & Imbens, 2016). The various adjustment parameters are also observed, such as minimum size of nodes for the trees and cross validation, minimizing the losses and the reduction of standard errors. The CF can be estimated using the causalTree package proposed by Athey (2019) for the R® software. See also the link to the code in Github (https://github.com/susanathey/causalTree). Other procedures and complements can be observed in the manual. We also suggest reading Vapnik (2000) for more information on ML.

3.4 Data and Variables Used

For the market, it would be interesting to identify companies before they present bankruptcy characteristics, minimizing investment losses. Such models or methodologies make the evaluation impartial, exempt from subjective influences, enabling the analyst to classify the risks of the company regarding its future and capacity to generate results.

For this verification, the bankruptcy prediction techniques are divided into: qualitative analysis, with subjective models; univariate analysis, using rates based on accounting data or market indicators; multivariate analysis, including discriminant analysis, logit, probit, non-linear, neural network, Altman z score, Ohlson o score, and models based on market value, among others (Altman & Hotchkiss, 2007). Models such as those of Altman (1968) use discriminant analysis to classify companies as solvent and insolvent.

Limitations of these studies are found when non-linear relationships may be presented between the variables studied, such as bankruptcy and the main company indicators (leverage, profitability, liquidity) (Giordani et al., 2014). Other limitations are of a modeling nature, such as the normality of the data used for the discriminant analysis, as well as the linearity of the variables. One problem associated with neural networks relates to the understanding and resolutions of the patterns found.

Regarding the causes of bankruptcy, there is no predominant isolated factor of company bankruptcy. The first studies used only endogenous variables, related to profitability, liquidity, and leverage indicators (Altman, 1968; Deakin, 1972; Ohlson, 1980). Following the same line with internal variables, Giordani et al. (2014) adopted the augmented standard logit methodology, in which they sought to understand the non-linear relationships of the variables that influence bankruptcy, and found significant and robust results.

In addition, there are the arguments that company bankruptcy suffers from an external influence, that is, exogenous variables related to the country's economic situation or to government policies, since the internal indicators do not present sufficient information about the economic conditions faced by companies (Johnson, 1970). Giordani et al. (2014) also suggest the inclusion of variables external to the bankruptcy models and also warn of the need for non-linear approaches.

Regarding the exogenous variables, there are arguments showing that smaller companies are more likely to fail due to various factors, such as: (i) bigger companies appear to more easily take advantage of the effects of scale; (ii) bigger companies have more bargaining power with suppliers and financial institutions, among others; and (iii) bigger companies tend to benefit from greater experience or learning (Strömberg, 2000).

It also warrants mentioning that, in some situations, it is advisable to build specific models for the sector, where there is a distinction between the size of the companies (Mensah, 1984; Taffler, 1984). A summary of some studies and variables can be observed in Table 2.

Table 2
Some variables used in the bankruptcy models

Endogenous variables	Authors
Net working capital/TA	Beaver (1966), Altman (1968), Deakin (1972), Altman, Haldeman, and Narayanan (1977)
Retained earnings/TA	Altman (1968), Altman et al. (1977), Ohlson (1980)
EBITDA/TA	Deakin (1972)
EBIT/TA	Altman (1968), Altman et al. (1977), Giordani et al. (2014)
Market value of NE/BVL	Altman (1968)
Sales/TA	Altman (1968)
Net rate/TA	Beaver (1966), Deakin (1972)
Total liabilities/TA	Beaver (1966), Deakin (1972), Ohlson (1980), DeYoung (2003), Jostarndt and Sautner (2010), Giordani et al. (2014)
Current assets/TA	Deakin (1972)
Working capital/TA	Deakin (1972), Ohlson (1980), Cole and Gunther (1995)
Cash/TA	Deakin (1972)
Cash flow/TA	Beaver (1966)
Current assets/CL	Beaver (1966), Deakin (1972), Altman et al. (1977), Ohlson (1980)
Liquid current assets/CL	Deakin (1972), Giordani et al. (2014)
Cash/Current liabilities	Deakin (1972)
Current assets/Sales	Deakin (1972)
Liquid current assets/Sales	Deakin (1972)
Cash/Sales	Deakin (1972)
Working capital/Sales	Deakin (1972)
Fund reserves/TA	Ohlson (1980)
Exogenous variables	
Size	Altman et al. (1977), Ohlson (1980), Cole and Gunther (1995), Strömberg (2000), DeYoung (2003), Jostarndt and Sautner (2010), Giordani et al. (2014)
Gross domestic product	Giordani et al. (2014)
Age	Jostarndt and Sautner (2010), Giordani et al. (2014)

TA = total assets; EBIT = earnings before interest and taxes; EBITDA = earnings before interest, taxes, depreciation, and amortization; NE = net equity; BVL = book value of liabilities; CL = current liabilities; Ln = natural logarithm.
Source: *Elaborated by the authors.*

Giordani et al. (2014) emphasize that the internal indicators are often explored in insolvency analyses, reflecting the capital structure, profitability, and liquidity of companies. In regard to leverage, the authors argue that, in bankruptcy conditions, liabilities exceed assets. Regarding profit and liquidity, these provide relevant information about the scarcity of liquid assets to give continuity to the company's activities, with continuous expenses and debt payment.

Low net working capital is a frequent problem presented by companies in bankruptcy situations, since resources are constantly consumed by the operating losses, reducing the proportion of current assets, generally represented by the company's liquidity. In regard to retained earnings/total assets, they indicate that newer companies tend to have lower earnings than companies consolidated in the market. According to Altman (1968), this individually tested variable was the most relevant for dividing the groups into bankrupt and non-bankrupt companies.

Debt structure is also relevant for explaining company bankruptcy. Companies that are more indebted with banks are more likely to restructure due to the greater ease of renegotiating their debt (Jostarndt & Sautner, 2010). The insolvency risk of big companies is reduced due to the large volume of assets, that is, they are too big to fail (Acharya & Mora, 2015), giving greater relevance to the Size variable. The inclusion of sector variables would make up for the economic variations caused by market oscillations, especially due to some financial, sector-related, technological, or supply-related crisis, among others.

4. DATA ANALYSIS

In recent years, the literature on ML has worked hard to produce quality estimates, even for large volumes of data. The predictions can be used to guide small populations with specific characteristics, such as corporate bankruptcy. With the aim of analyzing the heterogeneity among companies in the market, various accounting and sector-related variables for 1,247 companies were listed.

One thousand two hundred forty-seven U.S. companies were chosen from 10 sectors classified according to the Thomson Reuters Business Classification. The balance sheets chosen involve the five years of the bankruptcy process, as there is proof of declines in the indicators (Kalay, Singhal, & Tashjian, 2007). Among these companies, 66 filed for bankruptcy, 22 of which went bankrupt before 2008 and 44 after – the treatment period. For a closer measurement, the balance sheets of the non-bankrupt companies were collected in the same year as the bankrupt ones, totaling 32,188 quarterly observations retrieved from the Thomson Reuters database.

A large sample imbalance is perceived, with 1,181 non-bankrupt companies and 66 bankrupt ones, characterizing the unequal proportion between the two classes (bankrupt and non-bankrupt). To resolve this problem, the synthetic minority oversampling technique (SMOTE) methodology was used. The SMOTE is an algorithm for generating artificial data to balance the minority class based on the closest neighbors. The majority class is also resampled, increasing the volume of data (Chawla, Bowyer, Hall, & Kegelmeyer, 2002).

In regard to the variables, when applying the tree and CF methodology, as well as the other ML techniques, a greater number of variables would be interesting, with the aim of capturing the company characteristics in detail. This process generates considerable difficulty, as there are absent data in much of the balance sheets, thus compromising a high number of observations. We thus list a set of equity and sector variables in order to apply the methodology. The descriptive statistics without the synthetic data can be observed in Table 3.

Table 3
Descriptive statistics of the data

Name	Abbreviation	Mean	SD	Minimum	Median	3-quantiles	Maximum
Total equity	TE_I	0.38	3.06	-272.70	0.61	0.78	9.28
Total liabilities	TL_I	0.67	6.36	-0.09	0.39	0.60	730.98
Long-term liabilities	TLTD_I	0.13	0.42	0.00	0.02	0.16	25.97
Total net receivables	TRN_I	0.17	0.15	-0.21	0.14	0.23	10.81
Total revenue	TR_I	0.33	0.86	-6.58	0.27	0.41	134.18
Equipment	PPETN_I	0.22	0.23	-0.17	0.15	0.31	11.79
Retained earnings	RE_AD_I	-3.53	95.27	-16,217.69	-0.06	0.32	4.69
Total assets	LN_TA	17.76	1.78	5.01	17.90	18.98	33.96
Current assets	TCA_I	0.63	5.86	0.00	0.62	0.78	756.76
Current liabilities	TCL_I	0.43	2.93	0.00	0.22	0.36	273.70
Total debt	TD_I	0.24	1.14	0.00	0.08	0.27	77.76
Gross profit	GP_I	0.11	0.26	-8.50	0.09	0.14	32.79
Net income after tax	NIAT_I	-0.07	1.03	-76.59	0.00	0.02	15.27
Net sales	NS_I	0.33	0.86	-6.58	0.26	0.41	134.18
Short-term debts	NPSTD_I	0.05	0.65	0.00	0.00	0.00	59.14
Operating income	OI_I	-0.05	0.86	-76.82	0.01	0.03	15.82
Cost of products	CR_I	0.21	0.65	-5.69	0.15	0.27	101.39
EBIT	EBIT_I	-0.06	1.74	-175.93	0.01	0.03	15.82
EBITDA	EBITDA_I	-0.04	1.67	-165.73	0.02	0.04	15.82
Accounts payable	AP_I	0.13	0.81	0.00	0.06	0.12	92.98
Accrued expenses	AE_I	0.11	0.65	-21.24	0.06	0.10	73.80
Cash and equivalents	CSTI_I	0.23	0.23	-0.01	0.15	0.37	3.61
Stock	CST_I	0.23	0.23	-0.01	0.15	0.37	3.61
Sector dummies							
Technology	D_T	0.24	0.43	0.00	0.00	0.00	1.00

Table 3
Cont.

Name	Abbreviation	Mean	SD	Minimum	Median	3-quantiles	Maximum
Basic materials	D_BM	0.05	0.23	0.00	0.00	0.00	1.00
Cyclical consumption	D_CC	0.18	0.39	0.00	0.00	0.00	1.00
Non-cyclical consumption	D_CNC	0.05	0.22	0.00	0.00	0.00	1.00
Energy	D_E	0.06	0.25	0.00	0.00	0.00	1.00
Financial	D_F	0.02	0.14	0.00	0.00	0.00	1.00
Health	D_H	0.17	0.37	0.00	0.00	0.00	1.00
Industry	D_I	0.19	0.39	0.00	0.00	0.00	1.00
Telecommunications	D_TS	0.01	0.12	0.00	0.00	0.00	1.00
Utilities	D_U	0.01	0.08	0.00	0.00	0.00	1.00

Note: *Values in percentages. All the variables were weighted by total assets. For the total assets variable, the natural logarithm was used.*
SD = standard deviation; EBIT = earnings before interest and taxes; EBITDA = earnings before interest, taxes, depreciation, and amortization.
Source: *Elaborated by the authors.*

As expected, there is great variety among the companies, especially in size. This variation contributes substantially to the heterogeneity of the companies. It is also observed that despite there being many accounting variables, there is a low correlation between them (Figure 1).

The *X29* variable refers to the binary variable, indicating bankrupt or non-bankrupt firms, and the *TRA* variable refers to the binary treatment variable – before and after the crisis. It warrants mentioning that we are not interested in the causal effects caused based, especially, on parametric metrics, but in analyzing some variables that may indicate relevant partitions to indicate the soundness of a company. Within this context, the results of the CF cannot be interpreted as partial effects, keeping the other variables constant.

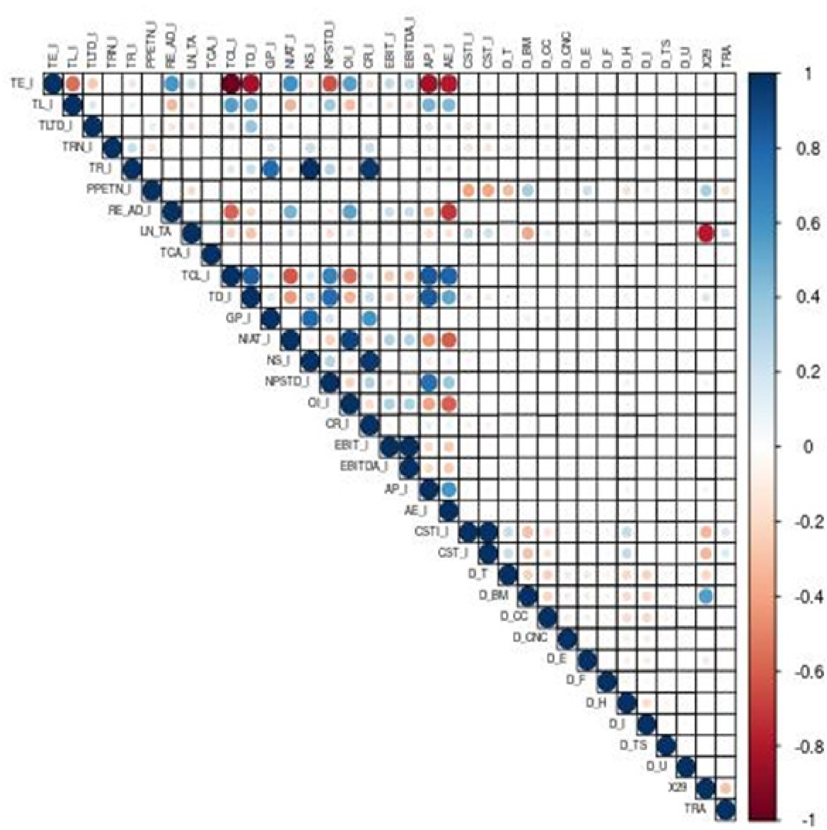

Figure 1 *Correlation between the variables*
Source: *Elaborated by the authors.*

4.1 Post-Lasso Analysis

A simple way of analyzing the causal effects between the pre- and post-financial collapse variables would be via simple interactions with a linear model, as described in equation 4. Athey and Imbens (2016) warn that this methodology would be relevant in models with few variables, becoming a problem when there is a large volume. With large sizes, one solution would be to carry out the Lasso as a kind of operator for choosing variables that are relevant to the model (Athey, Imbens, Pham, & Wager, 2017) and then applying the OLS regression (Belloni et al., 2014b). Having carried out these procedures, the results can be observed in Table 4.

Table 4
Post-Lasso results

Variables	Estimates	Standard error	Pr(>|t|)	Variables	Estimates	Standard error	Pr(>|t|)
(Intercept)	1.11451	0.00129	0.00000	I(TE_I * W)	0.11359	0.00176	0.00000
TL_I	0.00088	0.00010	0.00000	I(TRN_I * W)	-0.01963	0.00547	0.00033
TLTD_I	0.15660	0.00199	0.00000	I(PPETN_I * W)	0.06758	0.00502	0.00000
PPETN_I	0.21366	0.00350	0.00000	I(LN_TA * W)	-0.01647	0.00013	0.00000
RE_AD_I	0.00021	0.00001	0.00000	I(TCA_I * W)	0.00220	0.00023	0.00000
TCA_I	-0.00169	0.00015	0.00000	I(TCL_I * W)	0.12330	0.00187	0.00000
OI_I	0.00277	0.00163	0.09050	I(GP_I * W)	-0.05196	0.00370	0.00000
CR_I	-0.00059	0.00139	0.67372	I(NPSTD_I * W)	0.02562	0.00109	0.00000
EBITDA_I	0.00006	0.00048	0.90396	I(OI_I * W)	-0.01134	0.00189	0.00000
AP_I	-0.01409	0.00108	0.00000	I(AE_I * W)	-0.00835	0.00219	0.00013
D_BM	0.56272	0.00166	0.00000	I(CSTI_I * W)	0.12206	0.00382	0.00000
D_CC	0.11772	0.00148	0.00000				
D_CNC	0.13174	0.00248	0.00000				
D_TS	0.13751	0.00470	0.00000				

Lasso = least absolute shrinkage and selection operator.
Source: *Elaborated by the authors.*

With the interactions, the model would have 66 variables, of which 33 are the initial ones of the model (33 variables, 23 of which are accounting and 10 are sector-related) and 33 are interactions. It is observed that the volume of relevant interactions $I(*W)$, especially in the internal company variables, is high, totaling 11. The sector indicatives D were only relevant on four occasions, revealing that, before the financial crisis, the Basic materials (*D_BM*), Cyclical consumption (*D_CC*), Non-cyclical consumption (*D_CNC*), and Telecommunications (*D_TS*) sectors were the most affected in the bankruptcy processes. After the crisis, the results would be broad, with no relevant interactions. However, there is a limitation regarding the interpretation of this model, as it concerns a linear regression.

These results are very generic in terms of possible predictability, since different effects are found in a wide variety of companies. Given the individual characteristics of each company, the possibility of renegotiating debts, for example, would cause distortions regarding the possibilities of intervention in the companies. Another relevant point would be the characteristics of current assets in terms of the quick ratio and burn rate. The operating and non-operating income, as well as the quality of the earnings involved, may be relevant determinants for a company going bankrupt or not. And with these results (Table 4), the variables are treated homogenously.

4.2 Conditional Treatment and Causal Tree Analysis

In this context, there is the need to know in which subpopulations the financial crisis had the greatest effect. Athey and Imbens (2016) state that in these cases a data-oriented way of identifying the relevant heterogeneity may be convenient. Causal trees produce this indication based on the data in order to understand the heterogeneity and where it is according to the space of each variable, generating impartial estimates of the treatment in each subgroup. The initial tree was generated with 294 leaves. The cross-validation error (*x-val*) does not always reduce when the tree becomes more complex (to make it easy to understand, an analogy to the regression model is used: with the inclusion of more variables in the model, its predictive power does not increase). A good cut-off point would be when the points cut and are located below the horizontal line, opting for the point furthest to the left, generally the lowest *xerror* value. After all these analysis procedures, the regularization parameter converges in 156 divisions – the *xerror* value ceases to decrease.

It is also known that the interaction coefficients generated are the mean treatment effects of each one of the leaves (Table 5). After the adjustments, the tree would thus have 156 leaves. It is also known that in all these leaves the treatments are relevant.

Table 5
Effect of the treatment per leaf

Leaf	Estimate	Leaf	Estimate	Leaf	Estimate	Leaf	Estimate
Leaf_1	-1	Leaf_40	-0.89796	Leaf_79	-0.39655	Leaf_118	0.38961
Leaf_2	-0.99813	Leaf_41	-0.89305	Leaf_80	-0.34615	Leaf_119	0.39011
Leaf_3	-0.99448	Leaf_42	-0.89286	Leaf_81	-0.34211	Leaf_120	0.40705
Leaf_4	-0.99375	Leaf_43	-0.88889	Leaf_82	-0.3125	Leaf_121	0.45554
Leaf_5	-0.99058	Leaf_44	-0.88462	Leaf_83	-0.30303	Leaf_122	0.53782
Leaf_6	-0.98944	Leaf_45	-0.88413	Leaf_84	-0.30189	Leaf_123	0.54167
Leaf_7	-0.98936	Leaf_46	-0.875	Leaf_85	-0.29412	Leaf_124	0.56061
Leaf_8	-0.98924	Leaf_47	-0.87097	Leaf_86	-0.27778	Leaf_125	0.6
Leaf_9	-0.98221	Leaf_48	-0.86957	Leaf_87	-0.27451	Leaf_126	0.62372
Leaf_10	-0.98077	Leaf_49	-0.86538	Leaf_88	-0.2525	Leaf_127	0.63333
Leaf_11	-0.97857	Leaf_50	-0.86207	Leaf_89	-0.25	Leaf_128	0.66885
Leaf_12	-0.97619	Leaf_51	-0.86	Leaf_90	-0.22222	Leaf_129	0.74868
Leaf_13	-0.97464	Leaf_52	-0.85185	Leaf_91	-0.2151	Leaf_130	0.78481
Leaf_14	-0.97297	Leaf_53	-0.84783	Leaf_92	-0.14474	Leaf_131	0.79081
Leaf_15	-0.96985	Leaf_54	-0.84328	Leaf_93	-0.14444	Leaf_132	0.79094
Leaf_16	-0.9697	Leaf_55	-0.84	Leaf_94	-0.05294	Leaf_133	0.8
Leaf_17	-0.96769	Leaf_56	-0.83871	Leaf_95	-0.04819	Leaf_134	0.8125
Leaf_18	-0.96636	Leaf_57	-0.81731	Leaf_96	-0.04762	Leaf_135	0.81818
Leaf_19	-0.96592	Leaf_58	-0.78571	Leaf_97	-0.03509	Leaf_136	0.82467
Leaf_20	-0.96552	Leaf_59	-0.78182	Leaf_98	-0.02817	Leaf_137	0.82524
Leaf_21	-0.96226	Leaf_60	-0.75177	Leaf_99	-0.02542	Leaf_138	0.82927
Leaf_22	-0.96104	Leaf_61	-0.75	Leaf_100	-0.01471	Leaf_139	0.83888
Leaf_23	-0.95808	Leaf_62	-0.67701	Leaf_101	-0.01407	Leaf_140	0.84615
Leaf_24	-0.95288	Leaf_63	-0.67059	Leaf_102	-0.01316	Leaf_141	0.86194
Leaf_25	-0.94767	Leaf_64	-0.66667	Leaf_103	-0.00855	Leaf_142	0.86517
Leaf_26	-0.94643	Leaf_65	-0.66365	Leaf_104	-0.00131	Leaf_143	0.86842
Leaf_27	-0.9449	Leaf_66	-0.65625	Leaf_105	-0.00031	Leaf_144	0.89836
Leaf_28	-0.93023	Leaf_67	-0.65476	Leaf_106	-0.00136	Leaf_145	0.90398
Leaf_29	-0.9292	Leaf_68	-0.65385	Leaf_107	0.00201	Leaf_146	0.91525
Leaf_30	-0.92593	Leaf_69	-0.64706	Leaf_108	0.00678	Leaf_147	0.92011
Leaf_31	-0.92126	Leaf_70	-0.62805	Leaf_109	0.00797	Leaf_148	0.92537
Leaf_32	-0.92	Leaf_71	-0.6156	Leaf_110	0.01109	Leaf_149	0.94
Leaf_33	-0.91824	Leaf_72	-0.58696	Leaf_111	0.01667	Leaf_150	0.94231
Leaf_34	-0.91701	Leaf_73	-0.58283	Leaf_112	0.02041	Leaf_151	0.95187
Leaf_35	-0.91667	Leaf_74	-0.56604	Leaf_113	0.0303	Leaf_152	0.9575
Leaf_36	-0.91463	Leaf_75	-0.53061	Leaf_114	0.05085	Leaf_153	0.96364
Leaf_37	-0.91183	Leaf_76	-0.43836	Leaf_115	0.15094	Leaf_154	0.96923
Leaf_38	-0.9108	Leaf_77	-0.42	Leaf_116	0.24316	Leaf_155	0.975
Leaf_39	-0.90691	Leaf_78	-0.40789	Leaf_117	0.30556	Leaf_156	1

Note: *The standard error has 0.03762 and 0.00075 as its maximum and minimum values, respectively.*
Source: *Elaborated by the authors.*

The analyses are similar to an OLS regression. It is observed that the data are in decreasing order and only from leaf 107 onward are the coefficients positive; thus, the crisis would have a negative effect on more than half of the leaves, showing the relevance for the accounting variables analyzed.

Given the company conditions and their particularities, the financial crisis that occurred affected the various companies differently, since the effect of the treatment is different in each one of the leaves, calculated using the *F* test. It also warrants mentioning that if a division did not occur in a specific variable, it does not mean its irrelevance. There are various ways to choose a subsample with a wide variety of treatment effects, which can be high or low.

The general mean effect (mean of the variables) can be observed in Table 6. The sector variables, as highlighted, were the ones that presented a mean treatment close to 0 for the various leaves of the tree, indicating lower heterogeneity. Basic materials (*D_BM*) and Cyclical consumption (*D_CC*) stand out as the most affected sectors, having the most relevance at times of crisis, these being the most predominant sectors in terms of company bankruptcies after the crisis period.

Table 6
General mean per variable

Name	Abbreviation	Mean	Name	Abbreviation	Mean
Total equity	TE_I	-0.031596154	EBIT	EBIT_I	-0.068320513
Total liabilities	TL_I	1.085089744	EBITDA	EBITDA_I	-0.050916667
Long-term liabilities	TLTD_I	0.233576923	Accounts payable	AP_I	0.188685897
Total net receivables	TRN_I	0.14500641	Accrued expenses	AE_I	0.133032051
Total revenue	TR_I	0.388935897	Cash and equivalents	CSTI_I	0.120948718
Equipment	PPETN_I	0.309858974	Stock	CST_I	0.120948718
Retained earnings	RE_AD_I	-3.572378205	Technology	D_T	0.102634615
Total assets	LN_TA	15.31235897	Basic materials	D_BM	0.191153846
Current assets	TCA_I	0.488692308	Cyclical consumption	D_CC	0.312897436
Current liabilities	TCL_I	0.697269231	Non-cyclical consumption	D_CNC	0.097602564
Total debt	TD_I	0.518230769	Energy	D_E	0.029730769
Gross profit	GP_I	0.108211538	Financial	D_F	0.004371795
Net income after taxes	NIAT_I	-0.120929487	Health	D_H	0.091384615
Net sales	NS_I	0.389647436	Industry	D_I	0.148961538
Short-term debt	NPSTD_I	0.118884615	Telecommunications	D_TS	0.020955128
Operating income	OI_I	-0.094794872	Utilities	D_U	0.000269231
Cost of products	CR_I	0.281365385			

EBIT = earnings before interest and taxes; EBITDA = earnings before interest, taxes, depreciation, and amortization.
Source: *Elaborated by the authors.*

Companies that operate in sectors such as Utilities, Financial, Telecommunications, Energy, Health, Non-cyclical consumption, and Technology are the least affected by the financial crisis, possibly due to the need for the items produced. In regard to the variables used, it is observed that the most affected would be Net equity, EBITDA, EBIT, Operating income, Income after taxes, and Retained earnings. As expected, the Profit and Net equity variables had the negative effects with treatment means lower than 0, with retained earnings standing out with the lowest coefficient.

Due to the size of the estimated tree, which would be invisible in this document, it would not be possible to incorporate the figure, but the main segregation point would be the sector type the companies form part of. Standing out as a first division is the Basic materials (*D_BM*) sector and, for certain volumes in assets, smaller companies ($LN\_TA < e^{12.238}$), the next division would be Retained earnings. For companies that do not belong to the Basic materials sector (< 0.5), the next partition would be in Total Assets (*LN_TA*), where, for those bigger than LN_TA $e^{12.238}$, the segregation would be the Cyclical consumption

(*D_CC*) sector, highlighting that bigger companies tend to be less affected, presenting a high volume of subdivisions.

Characteristics such as Total liquid receivables (*TRN_I*) were shown to be relevant, given the need for an increase in company cash flows, especially at times of recession. Companies with *TRN_I*, for example, greater than 16% would tend to have bankruptcy points, depending on their size (*LN_TA*) and volume of debt (*TL_I*).

Not very far from what Giordani et al. (2014) presented, company size was relevant in the main partitions found, dampened by their high volume in assets, since smaller companies tend to be more prone to bankruptcy. There is also the possibility of more benefits and government interventions, aiming to dampen the amount of unemployment generated by large company bankruptcies.

Another important variable would be Net sales, converging with one of the indicators proposed by Altman (1968), showing that companies with more capacity to generate revenues present fewer problems in crisis periods. The liquidity variables were also relevant, as well as the profitability indicators.

4.3 CFs

CFs are therefore an adaptive and efficient method for estimating parameters that can be defined by local conditions, such as after applying the CATE. The predictions of the CF are mean causal tree estimates; that is, at least two causal trees are estimated and then the trees are combined, generating the CF estimates. The weights found in each one of the leaves of the causal trees reveal greater reliability in the volume of important dimensions, as well as being adaptive, making the estimates more robust in the face of company heterogeneity.

By predicting the CATE estimates and their variation for each observation, little variability is found, with a general mean close to 0 (Table 7) on the Predictions and Estimated variance lines. The term "Biased error," on the line, indicates that the error is only due to the variability of the data sample; that is, it represents the error that is expected with the construction of the forest containing an infinite number of trees. With this, the consistency of the estimates is noted, with an error close to 0.

Table 7
General mean of the conditional average treatment effect (CATE)

General mean of the CATE of the test sample							
	Mean	**SD**	**Minimum**	**1st quartile**	**Median**	**3rd quartile**	**Maximum**
Predictions	0	0.04	-1.48	0	0	0	0.68
Estimated variance	0	0.01	0.00	0	0	0	2.98
Biased error	0	0.00	0.00	0	0	0	0.27
General mean of the CATE of the validation sample							
Predictions	0	0.04	-1.28	0	0	0	0.68
Estimated variance	0	0.01	0.00	0	0	0	0.64

SD = standard deviation.
Source: *Elaborated by the authors.*

Based on the predictions of the test set, we estimated the predictions for the validation sample in Table 7. As expected, the estimates presented very small variations, all close to 0, indicating that the model fits the parameters and the data well. Therefore, the results converge toward a greater predictability possibility, as well as treating the characteristics of the companies analyzed homogenously. A reduction in the maximum value of the estimated variance is also found, reducing the previous threshold of 2.98 to 0.64. The Biased error variable does not appear, since it was tested in the validation sample.

The most used variables in the partition of the tree can be seen in Table 8. However, we cannot fall into the trap where, with little frequency of use in the partitions, the variable is not relevant. Observe that the frequency of the sector variable *D_BM* is 0.2%, but the main partition of the tree is found in that variable.

Table 8
Most used variables in the partition

Variable	Frequency	Variable	Frequency	Variable	Frequency
GP_I	0.26701	OI_I	0.01747	AE_I	0.00589
AP_I	0.14584	TL_I	0.01334	NS_I	0.00516
EBIT_I	0.09351	RE_AD_I	0.01318	D_BM	0.00255
EBITDA_I	0.07489	TCA_I	0.01260	D_I	0.00209
TLTD_I	0.05646	CR_I	0.00979	D_TS	0.00130
PPETN_I	0.04446	CST_I	0.00881	D_CC	0.00038
TE_I	0.04394	NPSTD_I	0.00849	D_T	0.00012
TRN_I	0.04311	CSTI_I	0.00780	D_CNC	0.00012
LN_TA	0.04266	TD_I	0.00709	D_E	0.00000
D_H	0.03818	TR_I	0.00678	D_F	0.00000
NIAT_I	0.02083	TCL_I	0.00614	D_U	0.00000

Source: *Elaborated by the authors.*

In the subpartitions, the Gross profit (GP_I) variable was the one that presented the highest frequency when the tree was divided, with approximately 27% of the appearances. The Accounts payable (AP_I) variable is relevant in the process of determining the bankruptcy of the companies, as it directly affects their cash flows, as well as their credibility. It also warrants mentioning that if two variables are highly correlated, there may be partitioning in one of the variables, but not in the other. However, if one is removed, the subdivision can occur in the one that was left, keeping the definitions in each leaf unaltered.

5. CONCLUDING REMARKS

The results indicated that there are several variables that are not normally included in the bankruptcy analysis and prediction models. The Net sales (*NS_I*) variable, according to Altman (1968), continues to be relevant. It warrants mentioning the importance of including variables that indicate the operating sector. It could be speculated that there are sectors that are more prone to bankruptcy, especially at times of crisis. In this research, the most affected was that of Basic materials (D_BM), which includes chemical, mineral exploration, and environmental (paper, wood, and recipients) companies. If it does not belong to *D_BM*, another highly affected sector would be Cyclical consumption (*D_CC*) (automobiles, construction material, domestic utensils, hotels, production, and entertainment).

We also observed the presence of heterogeneity among the companies, which in many cases are treated as identical. The debt ratios, for example, in linear models are treated as similar among the companies and they are not, given the size and bargaining capacity with suppliers and the government, among others.

Smaller-sized companies can also present less capacity for obtaining credit, requiring of managers larger amounts in cash or equivalents to remain functioning. With this, they tend to present higher liquidity indicators. Depending on the segment, companies can present greater amounts of fixed assets, reducing liquidity ratios; on the other hand they present larger volumes in depreciation. These characteristics should be taken into consideration in the treatment or intervention, especially in crisis periods, and it is up to the interventionists to adopt the best strategy for each company.

One limitation of this methodology would be the need for a quasi-experimental approach, requiring a database before and after a specific phenomenon. Analyzing without the need for this event would provide a greater academic contribution. It is suggested that future studies explore the unobserved characteristics of companies using other methodologies, addressing, for example, the intertemporal impact on companies and on the variables, as this proposed methodology would not address such effects and their magnitudes.

REFERENCES

Acharya, V. V., & Mora, N. (2015). A crisis of banks as liquidity providers. *Journal of Finance*, *70*(1), 1-43. https://doi.org/10.1111/jofi.12182

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

Altman, E. I., Haldeman, R. G., & Narayanan, P. (1977). ZETA analysis: A new model to identify bankruptcy risk of corporations. *Journal of Banking and Finance*, *1*(1), 29–54. https://doi.org/10.1016/0378-4266(77)90017-6

Altman, E. I., & Hotchkiss, E. (2007). *Corporate financial distress and bankruptcy* (3a ed.). Hoboken, NJ: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118267806

Antunes, F., Ribeiro, B., & Pereira, F. (2017). Probabilistic modeling and visualization for bankruptcy prediction. *Applied Soft Computing Journal*, *60*, 831-843. https://doi.org/10.1016/j.asoc.2017.06.043

Athey, S. (2015). Machine learning and causal inference for policy evaluation. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD'15* (p. 5-6). New York, NY. https://doi.org/10.1145/2783258.2785466

Athey, S. (2019). *CausalTree*. Retrieved from https://github.com/susanathey/causalTree

Athey, S., & Imbens, G. (2015). Machine learning methods in economics and econometrics: A measure of robustness to misspecification. *American Economic Review*, *105*(5), 476-480. https://doi.org/10.1257/aer.p20151020

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353-7360. https://doi.org/10.1073/pnas.1510489113

Athey, S., Imbens, G., Pham, T., & Wager, S. (2017). Estimating average treatment effects: Supplementary analyses and remaining challenges. *American Economic Review*, *107*(5), 278-281. https://doi.org/10.1257/aer.p20171042

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148-1178. https://doi.org/10.1214/18-AOS1709

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*, *4*, 71–111. https://doi.org/10.2307/2490171

Belloni, A., Chernozhukov, V., & Hansen, C. (2014a). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, *28*(2), 29-50. https://doi.org/10.1257/jep.28.2.29

Belloni, A., Chernozhukov, V., & Hansen, C. (2014b). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, *81*(2), 608-650. https://doi.org/10.1093/restud/rdt044

Benmelech, E., & Bergman, N. K. (2011). Bankruptcy and the collateral channel. *Journal of Finance*, *66*(2), 337-378. https://doi.org/10.1111/j.1540-6261.2010.01636.x

Boot, A. W. A., & Thakor, A. V. (1997). Financial system architecture. *Review of Financial Studies*, *10*(3), 693-733. https://doi.org/10.1093/rfs/10.3.693

Brogaard, J., Li, D., & Xia, Y. (2017). Stock liquidity and default risk. *Journal of Financial Economics*, *124*(3), 486-502. https://doi.org/10.1016/j.jfineco.2017.03.003

Chauhan, N., Ravi, V., & Chandra, D. K. (2009). Differential evolution trained wavelet neural networks: Application to bankruptcy prediction in banks. *Expert Systems With Applications*, *36*(4), 7659-7665. https://doi.org/10.1016/j.eswa.2008.09.019

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*(1), 321-357. https://doi.org/10.1613/jair.953

Cho, S., Kim, J., & Bae, J. K. (2009). An integrative model with subject weight based on neural network learning for bankruptcy prediction. *Expert Systems With Applications*, *36*(1), 403-410. https://doi.org/10.1016/j.eswa.2007.09.060

Cielen, A., Peeters, L., & Vanhoof, K. (2004). Bankruptcy prediction using a data envelopment analysis. *European Journal of Operational Research*, *154*(2), 526-532. https://doi.org/10.1016/S0377-2217(03)00186-3

Cole, R. A., & Gunther, J. W. (1995). Separating the likelihood and timing of bank failure. *Journal of Banking and Finance*, *19*(6), 1073–1089. https://doi.org/10.1016/0378-4266(95)98952-M

Deakin, E. B. (1972). A discriminant analysis of predictors of business failure. *Journal of Accountin Research*, *10*(1), 167-179. Retrieved from http://www.jstor.org/stable/2490225

DeMarzo, P. M., & Fishman, M. J. (2007). Optimal long-term financial contracting. *Review of Financial Studies*, *20*(6), 2079-2128. https://doi.org/10.1093/rfs/hhm031

DeSpiegeleer, J., Madan, D. B., Reyners, S., & Schoutens, W. (2018). Machine learning for quantitative finance: Fast derivative pricing, hedging and fitting. *Quantitative Finance*, *18*(10), 1635-1643. https://doi.org/10.1080/14697688.2018.1495335

DeYoung, R. (2003). The failure of new entrants in commercial banking markets: A split-population duration analysis. *Review of Financial Economics*, *12*(1), 7–33. https://doi.org/10.1016/S1058-3300(03)00004-1

FitzPatrick, P. J. (1932). *A comparison of the ratios of successful industrial enterprises with those of failed companies.* Retrieved from https://www.worldcat.org/title/comparison-of-the-ratios-of-successful-industrial-enterprises-with-those-of-failed-companies/oclc/6284198

García, V., Marqués, A. I., Sánchez, J. S., & Ochoa-Domínguez, H. J. (2017). Dissimilarity-based linear models for corporate bankruptcy prediction. *Computational Economics*, *53*, 1019-1031. https://doi.org/10.1007/s10614-017-9783-4

Giordani, P., Jacobson, T., Schedvin, E. Von, & Villani, M. (2014). Taking the Twists into account: Predicting firm bankruptcy risk with splines of financial ratios. *Journal of Financial and Quantitative Analysis*, *49*(4), 1071-1099. https://doi.org/10.1017/S0022109014000623

Helwege, J., & Zhang, G. (2016). Financial firm bankruptcy and contagion. *Review of Finance*, *20*(4), 1321-1362. https://doi.org/10.1093/rof/rfv045

Hertzel, M. G., Li, Z., Officer, M. S., & Rodgers, K. J. (2008). Inter-firm linkages and the wealth effects of financial distress along the supply chain. *Journal of Financial Economics*, *87*(2), 374-387. https://doi.org/10.1016/j.jfineco.2007.01.005

Hertzel, M. G., & Officer, M. S. (2012). Industry contagion in loan spreads. *Journal of Financial Economics*, *103*(3), 493-506. https://doi.org/10.1016/j.jfineco.2011.10.012

Hochreiter, S., & Schimidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Ivashina, V., Iverson, B., & Smith, D. C. (2016). The ownership and trading of debt claims in Chapter 11 restructurings. *Journal of Financial Economics*, *119*(2), 316-335. https://doi.org/10.1016/j.jfineco.2015.09.002

Johnson, C. G. (1970). Ratio Stability and corporate failure. *The Journal of Finance*, *25*(5), 1166-1168. https://doi.org/10.2307/2325590

Jorion, P., & Zhang, G. (2007). Good and bad credit contagion: Evidence from credit default swaps. *Journal of Financial Economics*, *84*(3), 860-883. https://doi.org/10.1016/j.jfineco.2006.06.001

Jostarndt, P., & Sautner, Z. (2010). Out-of-court restructuring versus formal bankruptcy in a non-interventionist bankruptcy setting. *Review of Finance*, *14*(4), 623-668. https://doi.org/10.1093/rof/rfp022

Kalay, A., Singhal, R., & Tashjian, E. (2007). Is Chapter 11 costly? *Journal of Financial Economics*, *84*(3), 772-796. https://doi.org/10.1016/j.jfineco.2006.04.001

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, *34*(11), 2767-2787. https://doi.org/10.1016/j.jbankfin.2010.06.001

Lang, L. H. P., & Stulz, R. (1992). Contagion and competitive intra-industry effects of bankruptcy announcements. An empirical analysis. *Journal of Financial Economics*, *32*(1), 45-60. https://doi.org/10.1016/0304-405X(92)90024-R

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

Lennox, C. (1999). Identifying failing companies: A re-evaluation of the logit, probit and DA approaches. *Journal of Economics and Business*, *51*, 347-364.

Ludwig, R. S., & Piovoso, M. J. (2005). A comparison of machine-learning classifiers for selecting money managers. *Intelligent Systems in Accounting, Finance and Management*, *13*(3), 151-164. https://doi.org/10.1002/isaf.262

Mensah, Y. M. (1984). An examination of the stationarity of multivariate bankruptcy prediction models: A methodological study. *Journal of Accounting Research*, *22*(1), 380. https://doi.org/10.2307/2490719

Min, J. H., & Lee, Y. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications*, *28*(4), 603-614. https://doi.org/10.1016/j.eswa.2004.12.008

Montenegro, M. R., & Albuquerque, P. H. M. (2017). Wealth management: Modeling the nonlinear dependence. *Algorithmic Finance*, *6*(1-2), 51-65. https://doi.org/10.3233/AF-170203

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

Park, C. (2000). Monitoring and structure of debt contracts. *The Journal of Finance*, *55*(5), 2157-2195. https://doi.org/10.1111/0022-1082.00283

Pendharkar, P. C. (2005). A threshold-varying artificial neural network approach for classification and its application to bankruptcy prediction problem. *Computers & Operations Research*, *32*(10), 2561-2582. https://doi.org/10.1016/j.cor.2004.06.023

Premachandra, I. M., Bhabra, G. S., & Sueyoshi, T. (2009). DEA as a tool for bankruptcy assessment: A comparative study with logistic regression technique. *European Journal of Operational Research*, *193*(2), 412-424. https://doi.org/10.1016/j.ejor.2007.11.036

Premachandra, I. M., Chen, Y., & Watson, J. (2011). DEA as a tool for predicting corporate failure and success: A case of bankruptcy assessment. *Omega*, *39*(6), 620-626. https://doi.org/10.1016/j.omega.2011.01.002

Rodano, G., Serrano-Velarde, N., & Tarantino, E. (2016). Bankruptcy law and bank financing. *Journal of Financial Economics*, *120*(2), 363-382. https://doi.org/10.1016/j.jfineco.2016.01.016

Strömberg, P. (2000). Conflicts of interest and market illiquidity in bankruptcy auctions: Theory and tests. *Journal of Finance*, *55*(6), 2641-2692. https://doi.org/10.1111/0022-1082.00302

Taffler, R. J. (1984). Empirical models for the monitoring of UK corporations. *Journal of Banking and Finance*, *8*(2), 199-227. https://doi.org/10.1016/0378-4266(84)90004-9

Tsai, C. F., Hsu, Y. F., & Yen, D. C. (2014). A comparative study of classifier ensembles for bankruptcy prediction. *Applied Soft Computing Journal*, *24*, 977-984. https://doi.org/10.1016/j.asoc.2014.08.047

Tseng, F., & Hu, Y. (2010). Comparing four bankruptcy prediction models: Logit, quadratic interval logit, neural and fuzzy neural networks. *Expert Systems With Applications*, *37*(3), 1846-1853. https://doi.org/10.1016/j.eswa.2009.07.081

Vapnik, V. N. (2000). *The nature of statistical learning theory* (2a ed.). New York, NY: Springer-Verlag. https://doi.org/10.1007/978-1-4757-3264-1

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3-28. https://doi.org/10.1257/jep.28.2.3

Varian, H. R. (2016). Intelligent technology. *Finance & Development*, *53*(3), 6-9.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228-1242. https://doi.org/10.1080/01621459.2017.1319839

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

Yaohao, P., & Albuquerque, P. H. M. (2019). Non-linear interactions and exchange rate prediction: Empirical evidence using support vector regression. *Applied Mathematical Finance*, *26*(1), 69-100. https://doi.org/10.1080/1350486X.2019.1593866

Yaohao, P., Albuquerque, P. H. M., Camboim de Sá, J. M., Padula, A. J. A., & Montenegro, M. R. (2018). The best of two worlds: Forecasting high frequency volatility for cryptocurrencies and traditional currencies with support vector regression. *Expert Systems with Applications*, *97*, 177-192. https://doi.org/10.1016/j.eswa.2017.12.004

Zhang, M., & Zhou, Z. (2007). ML-KNN : A lazy learning approach to multi-label learning. *Pattern Recognition*, *40*(7), 2038-2048. https://doi.org/10.1016/j.patcog.2006.12.019

Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research*, *22*, 83-86. https://doi.org/10.2307/2490860