ISSN 1983-1838
(DOI): 10.18624/etech.1342

eTECH TECNOLOGIAS PARA COMPETITIVIDADE INDUSTRIAL

PREDICTING INSOLVENCY IN BRAZILIAN COMPANIES USING MACHINE LEARNING: An Analysis of the Random Forest Technique

PREVISÃO DE INSOLVÊNCIA EM EMPRESAS BRASILEIRAS COM O USO DE MACHINE LEARNING: uma análise da técnica Random Forest

MICHELE APARECIDA CUNHA
https://orcid.org/0000-0003-1546-0504/ micheleapcunha@hotmail.com
Universidade Federal de Uberlândia – UFU, Uberlândia, Minas Gerais

ANTONIO SÉRGIO TORRES PENEDO
https://orcid.org/0000-0001-7763-8457/ penedo@ufu.br
Universidade Federal de Uberlândia – UFU, Uberlândia, Minas Gerais

FLÁVIO LUIZ DE MORAES BARBOZA
https://orcid.org/ 0000-0002-3449-5297/ flmbarboza@ufu.br
Universidade Federal de Uberlândia – UFU, Uberlândia, Minas Gerais

Recebido em: 11/06/2024.
Aprovado em: 05/11/2024.
Publicado em: 07/11/2024.

ABSTRACT

Corporate insolvency prediction serves as a valuable strategic parameter to proactively identify financial and managerial risks. Recent approaches to bankruptcy prediction have predominantly leveraged machine learning algorithms, demonstrating superior predictive accuracy compared to conventional methods. This study highlights the effectiveness of the Random Forest methodology in insolvency assessments. Consequently, an exploratory analysis was conducted to evaluate the feasibility of employing a bankruptcy prediction model within the scope of publicly traded Brazilian companies using machine learning techniques. The sample is made up of companies with delisting status on the B3 stock exchange due to insolvency, between 2005 and 2018, a period in line with the enactment of Brazilian legislation on the subject. Evaluation of the model in Brazilian companies revealed an accuracy of 98% in predicting bankruptcies, highlighting the effectiveness of the Random Forest model as a valuable resource for investors, corporate decision makers and interested parties. This research contributes significantly to the discourse surrounding the adoption of machine learning tools in the field of bankruptcy prediction in the Brazilian business scenario.

Keywords: Machine Learning; Prediction of Bankruptcy; Random Forest; Risk management.

ISSN 1983-1838
(DOI): 10.18624/etech.1342

eTECH
TECNOLOGIAS PARA
COMPETITIVIDADE INDUSTRIAL

PREDICTING INSOLVENCY IN BRAZILIAN COMPANIES USING MACHINE LEARNING: An Analysis of the Random Forest Technique

Michele Aparecida Cunha
Antonio Sérgio Torres Penedo
Flávio Luiz De Moraes Barboza

RESUMO

A previsão da insolvência corporativa serve como um parâmetro estratégico valioso para identificar proativamente os riscos financeiros e gerenciais. Abordagens recentes na previsão de falências alavancaram predominantemente os algoritmos de aprendizado de máquina, demonstrando uma precisão preditiva superior em comparação aos métodos convencionais. Este estudo destaca a eficácia da metodologia Random Forest nas avaliações de insolvência. Consequentemente, uma análise exploratória foi conduzida para avaliar a viabilidade de empregar um modelo de previsão de falências no âmbito das empresas brasileiras de capital aberto por meio de técnicas de aprendizado de máquina. A amostra é composta por empresas com status de deslistagem na bolsa B3 devido à insolvência, entre 2005 e 2018, período alinhado à promulgação da legislação brasileira sobre o assunto. A avaliação do modelo em empresas brasileiras revelou uma precisão de 98% na previsão de falências, ressaltando a eficácia do modelo Random Forest como um recurso valioso para investidores, tomadores de decisão corporativos e partes interessadas. Esta pesquisa contribui significativamente para o discurso em torno da adoção de ferramentas de aprendizado de máquina no campo da previsão de falências no cenário empresarial brasileiro.

Palavras-Chave: Aprendizado de Máquina; Gestão de Riscos; Previsão de Falência; Random Forest.

1 INTRODUCTION

The ability to anticipate financial risks is a significant protective shield against crises. As Teixeira (2015) metaphorically describes, a crisis is like a monster that devours everyone. In the business world, the repercussions of an organization's insolvency extend to various stakeholders, including entrepreneurs, shareholders, employees, suppliers, customers, governmental entities, and society at large. Amid uncertainties and economic instabilities, increasing challenges, market dynamics, diversification trends, the advent of new business models, and many other factors surrounding companies, the use of effective monitoring tools becomes imperative to assess the vitality and fiscal health of businesses.

It is worth noting an additional comparison found in the symbolic representations of Mercury's Caduceus and Asclepius' Rod (Prates, 2002). These symbols are respectively associated with economic activity and medical activity and share similarities in their representations. When their meanings are reproduced in the business context, these symbols suggest a connection between an organization's well-being and its financial oversight.

From this perspective, the objective of the present study is to examine a financial monitoring model to predict insolvencies in the Brazilian market using current methodologies based on machine learning or artificial intelligence models. These models facilitate predictive analysis by leveraging substantial datasets and causal hypotheses (Kohavi & Provost, 1998).

These models are widely used daily by platforms such as Netflix (Sigiliano & Faustino, 2016), Amazon, and Google (Shokri, et al. 2017) to provide users with recommendations on movies, books, and other content based on their previous selections. They are also applied in social media platforms like X (formerly Twitter) (Davidov, Tsur & Rappoport, 2010) for sentiment analysis, trending topics, and fraud detection. Furthermore, their application has been widespread in developments of the Internet of Things (IoT) and precision medicine (Chiavegatto Filho, 2015). Organizations dealing with substantial datasets use these models to identify opportunities, vulnerabilities, and fraudulent activities. The use of machine learning is a growing trend that contributes to better organizational effectiveness.

With the wide use of machine learning models in various domains, the investigation driving this study revolves around its potential application in finance research, particularly for bankruptcy prediction. The main research question guiding our investigation is: To what extent can machine learning methodologies accurately predict the bankruptcy of companies?

This research employed machine learning with a specific focus on leveraging the Random Forest technique to predict bankruptcies in publicly traded Brazilian companies that had their listing suspended on the Brazilian stock exchange, B3, due to insolvency.

The model presented by Barboza, Kimura, and Altman (2017) suggests that the Random Forest technique demonstrates superior predictive accuracy compared to conventional bankruptcy prediction models when implemented in American companies. Therefore, this technique was selected for application in this research, allowing an analysis of the model in the Brazilian context. The importance of the model lies in its direct, pragmatic, and consistent presentation.

This study seeks to evaluate the feasibility of employing the bankruptcy prediction model advocated by Barboza, Kimura, and Altman (2017) using the Random Forest technique in a sample of publicly traded Brazilian companies that went through bankruptcy or judicial recovery processes between 2005 and 2018. Consequently, it aims to contribute to the validation of a model applicable to the Brazilian market, which may be useful for the stakeholders of organizations.

The work is divided into four distinct sections. The initial section outlines the fundamental principles of Machine Learning, Random Forest, and Bankruptcy Prediction. These concepts were elucidated through an examination of relevant literature from prominent digital repositories such as IEEEXplore, Elsevier, Science Direct, and Scielo, using specific keywords pertinent to the thematic area.

The subsequent section describes the methodology and dataset used in the research. The adopted methodology was based on the Random Forest technique, which was selected due to its higher predictive accuracy compared to alternative methodologies. The framework initially proposed by Barboza, Kimura, and Altman (2017) was adapted to fit the Brazilian context and the availability of information. Data collection was carried out using the Economatica tool and data from the Brazilian Securities and Exchange Commission (CVM) of companies with delisted statuses between 2005 and 2018.

In the third section, the results derived from the implementation of the Random Forest technique are presented and discussed. The results suggest the utility of the model in the Brazilian context, presenting an accuracy of 98%. Additionally, the results draw a contrast illustrating the superior accuracy of the Random Forest approach compared to conventional models such as logistic regression, thus deserving the attention of economic and academic agents. Subsequently, the fourth section is dedicated to outlining the study's conclusions, delineating the constraints, and proposing avenues for future research.

2 REFERENTIAL THEORETICAL

2.1 Machine Learning

During the World Wars era, there was a remarkable increase in interest in analytical machines due to their encoding and decoding capabilities. The 1940s witnessed the emergence of machines capable of performing complex calculations autonomously. Simultaneously, the foundations for computing theory were being established. The renowned English mathematician Alan Turing made significant advances in 1950 through his publication titled "Can Machines Think?" (Turing, 1950), being considered the precursor in the field of Artificial Intelligence (Tasinaffo, 2008).

In 1956, the Dartmouth Conference was held in the United States, where the formalization of Artificial Intelligence as a science was advocated. Additionally, in 1959, Arthur Samuel presented the concept of machine learning, defined as the field of study dedicated to enabling the learning of computational systems through training that eventually eliminates the need for complex programming (Samuel, 1959). Since then, research on artificial intelligence has delved deeper into exploring practical applications of data in various sectors such as engineering, robotics, pattern recognition, among others (Cano et al., 2017).

The focus of this study is on the use of machine learning methodologies for predicting organizational insolvency. Several techniques have been employed previously for this analysis in the American context, such as linear discriminant analysis, logistic regression, neural networks, Support Vector Machine (SVM), linear SVM, radial boosting, bagging, and Random Forest (Barboza, Kimura, & Altman, 2017). The emphasis of this investigation is on the Random Forest technique attributed to its superior predictive accuracy for bankruptcies compared to other models. Furthermore, it seeks to verify its accuracy in an emerging economy, specifically the Brazilian context.

2.2 Random Forest

The Random Forest technique proposed by Breiman (2001) represents a sophisticated machine learning approach based on classification from decision tree structures that use independent variables uniformly distributed across various data samples. Each tree in this ensemble method contributes by voting for the most prevalent class, resulting in more accurate predictions.

This model operates through supervised learning, leveraging historical behaviors of variables during the training phase to make global predictions by averaging the results of individual decision trees. As emphasized by Boot et al. (2014), the ensemble of the random forest can be interpreted as a comprehensible human decision tree, thus enhancing its applicability in real-world situations.

The Random Forest model presents fast computational speed, ease of implementation, and good performance (Kruppa et al., 2013). Additionally, it offers the advantage of versatility in classification tasks and can be applied in logistic regression techniques (Cano et al., 2017), which supports various purposes. It also has a convenient ranking in terms of accuracy and efficiency (Calderoni et al., 2015).

Several studies demonstrate the superior predictive accuracy achieved by the Random Forest technique compared to conventional methods such as logistic regression and adjusted logistic regression, as well as other machine learning approaches like Support Vector Machine (SVM) and neural networks (Booth, Gerding, & Mcgroarty, 2014; Cano et al., 2017; Kruppa et al., 2013; Yeh, Chi, & Lin, 2014; Barboza, Kimura, & Altman, 2017).

The application of the Random Forest model is comprehensive and diverse. Calderoni et al. (2015) employed the model to develop algorithms for locating patients in hospital facilities. Cano et al. (2017) applied the technique in computational chemistry to perform virtual screening in the identification of molecules for drug discovery. Maione et al. (2015) used the model to distinguish the main components of organic versus non-organic juice.

In the financial sector, the application of Random Forest serves various purposes. Kruppa et al. (2013) used the technique to assess the risk of default associated with credit transactions. Booth et al. (2014) employed the model to predict trends in stock trading returns during seasonal occurrences. Xiao (2016) used the technique for detecting fraud in financial transactions. Yeh, Chi, and Lin (2014) used the technique to assess credit risk in Taiwanese financial institutions, and Barboza, Kimura, and Altman (2017) used the model to predict corporate insolvencies in publicly traded American companies.

Given its versatility, ease of implementation, and promising accuracy results compared to traditional models, the Random Forest technique was chosen for this study, allowing its assessment in the context of bankruptcy prediction for publicly traded Brazilian companies.Figure 1 – Title (do not use a period)

2.2 Bankruptcy Prediction

The initial research in bankruptcy prediction dates to the 1960s, with Beaver (1966) being one of the pioneers in the field. His study utilized financial ratios to predict bankruptcies, serving as a benchmark for future research. Later, Altman (1968) developed the renowned Z-Score model using multiple discriminant analysis to predict bankruptcies, a model that remains influential in academic literature and practical applications.

Bankruptcy prediction models aim to identify the likelihood of a company's insolvency based on its financial health. These models traditionally rely on financial ratios and statistical techniques to determine the probability of bankruptcy. However, recent advancements in machine learning have introduced new methodologies that enhance predictive accuracy and offer more robust solutions for bankruptcy prediction.

Machine learning techniques, such as the Random Forest model, provide a data-driven approach to bankruptcy prediction. These techniques leverage vast amounts of data and complex algorithms to identify patterns and trends that may indicate financial distress. By using historical data and training the model on known outcomes, machine learning models can predict the probability of bankruptcy with high accuracy.

The application of machine learning in bankruptcy prediction has gained traction due to its ability to handle large datasets, its flexibility in incorporating various variables, and its superior predictive performance compared to traditional statistical methods. This study aims to evaluate the effectiveness of the Random Forest technique in predicting bankruptcies in the Brazilian market, providing valuable insights for investors, corporate decision-makers, and other stakeholders.

3 METHODOLOGY

The methodology of the present study was based on applied research, with the objective of generating knowledge for practical application. Employing the quantitative method, open data from the financial statements of companies listed on the Brazilian stock exchange (B3) were utilized for an exploratory and descriptive study. Since the accounting records maintained by companies reflect the decisions made by their managers, serving as a significant indicator of their trajectory, they are increasingly vital in bankruptcy prediction models (Guimarães & Moreira, 2008). The research procedure was experimental, based on the bankruptcy prediction model proposed by Barboza, Kimura, and Altman (2017).

Based on Barboza, Kimura, and Altman (2017), the present study incorporates predictive variables from Altman's seminal model (1968), the Z-score model, and Carton and Hofer's (2006) organizational performance model. The Z-score model determines five relevant financial dimensions: liquidity (X1), profitability (X2), productivity (X3), leverage (X4), and asset turnover (X5).

Liquidity (Working capital/Total assets): measured as the difference between current assets and current liabilities. Altman (1968) considers this the best indicator of an organization's discontinuation.

Profitability (Retained earnings/Total assets): in this variable, the company's age is an influencing factor, as a recent company has not yet retained earnings, which is also one of the reasons why new companies have a relatively higher insolvency rate than older companies (Altman, 1968).

Productivity (Earnings before interest and taxes/Total assets): a measure of the productivity of the organization's assets (Altman, 1968).

Leverage (Market value of equity * number of shares/Total debt): this measure demonstrates how much the assets can decrease in value before liabilities exceed assets, causing the company to become insolvent (Altman, 1968).

Asset Turnover (Sales/Total assets): a measure of the organization's ability to cope with competitive conditions (Altman, 1968).

Barboza, Kimura, and Altman's (2017) model adds other indicators to the Z-Score model that provide visibility into short-term financial performance, such as asset growth (AG), sales growth (SG), growth in the number of employees, operating margin (OM), change in return on equity (CROE), and change in price/book value ratio (PBR).

For the present study, Barboza, Kimura, and Altman's (2017) model was adapted to the Brazilian context, omitting the variable of employee growth due to the lack of available information. The variables and their corresponding formulas are detailed in Table 1.

Table 1 – Predictive Variables

Variable	Code	Metric	Authors
Liquidity	X1	Working capital / Total assets	Altman (1968); Barboza, Kimura and Altman (2017)
Profitability	X2	Retained earnings / Total assets	Altman (1968); Barboza, Kimura and Altman (2017)
Productivity	X3	Earnings before interest and taxes / Total assets	Altman (1968); Barboza, Kimura and Altman (2017)
Leverage	X4	Market value of equity * number of shares / Total debt	Altman (1968); Barboza, Kimura and Altman (2017)
Asset Turnover	X5	Sales / Total assets	Altman (1968); Barboza, Kimura and Altman (2017)
Asset Growth	CA	Earnings before interest and taxes / Sales	Barboza, Kimura and Altman (2017)
Sales Growth	CV	Total assets t – Total assets t−1 / Total assets t−1	Barboza, Kimura and Altman (2017)
Operating Margin	MO	Sales t – Sales t−1 / Sales t−1	Barboza, Kimura and Altman (2017)
Change in Return on Equity	VROE	ROE t – ROE t−1, where ROE = Net income / Shareholders' equity	Barboza, Kimura and Altman (2017)
Change in Price/Book Value Ratio	VPB	Market value per book value t – Market value per book value t−1, where M/B = Market value per share / Book value per share	Barboza, Kimura and Altman (2017)

Source: Study data.

The research sample consisted of publicly traded companies whose listing was delisted from B3 between 2005 and 2018. Publicly traded companies were selected based on data accessibility and their economic significance. The selection of the timeframe was determined based on the enactment of Law No. 11.101 in 2005, which regulates judicial and extrajudicial recovery and bankruptcy, as well as data availability.

The database survey involved examining documents from publicly traded companies registered with the Brazilian Securities and Exchange Commission - CVM, and, like the studies by Santos et al. (2018), companies undergoing judicial or extrajudicial recovery, bankrupt companies, or companies that have been recovered were considered insolvent. Table 2 presents the list of companies included in the sample.

Table 2 – Companies in the research sample

Company Name	Sector	Recovery Petition	Final Judicial Recovery	Status
Bombril Holding S. A.	Diversified Holding	11/07/2005	10/15/2010	R
Buettner Sa Ind e Comércio	Textile and Apparel	05/05/2011	04/28/2016	B
CELPA - Centrais Elétricas do Pará S. A.	Electric Power	02/28/2012	12/02/2014	R
Cerâmica Chiarelli S. A.	Construction and Decoration Materials	12/30/2008		UR
Cia Indl Schlosser S. A.	Textile and Apparel	04/04/2011		UR
Const Sultepa S. A.	Construction and Decoration Materials	07/06/2015		UR
Construtora Beter S. A.	Construction and Decoration Materials	09/12/2008	02/21/2017	B
Eneva S. A.	Electric Power	12/09/2014	06/26/2016	R
Eucatex S. A. Ind e Comércio	Construction and Decoration Materials	10/30/2005	11/06/2009	R
Fáb. Tec. Carlos Renaux S. A.	Textile and Apparel	12/09/2011	07/15/2013	B
Fição e Tecelagem São Jose S/A	Textile and Apparel	07/12/2010		UR
Fibam Cia Industrial	Basic Materials	10/14/2014		UR
GPC Participações S. A.	Petrochemicals	04/09/2013		UR
IGB Eletrônica S. A.	Machinery and Industrial Equipment	04/23/2010		UR
Inepar Equip. e Mont. S. A.	Machinery and Industrial Equipment	08/29/2014		UR
Lark S. A. Maq e Equip.	Machinery and Industrial Equipment	06/05/2012	09/26/2013	B
Lupatech S. A.	Machinery and Industrial Equipment	05/26/2015		UR
Mangels Industrial S. A.	Metallurgy/Steel Industry	11/01/2013	03/14/2017	R
Metalurgica Duque Sa	Metallurgy/Steel Industry	02/03/2014	07/20/2015	B
MMX Miner. e Metálicos S. A.	Mineral Extraction	11/25/2016		UR
OGX	Oil and Gas	10/30/2013	08/02/2017	R
Oi S.A.	Telecommunications	06/20/2016		UR
OSX Brasil S.A.	Equipment and Services	11/11/2013		UR
Recrusul S.A.	Machinery and Industrial Equipment	01/25/2006	12/23/2008	R

Rede Energia Particip. S. A.	Electric Power	11/23/2012	08/31/2016	R
Refinaria Manguinhos S. A.	Oil and Gas	01/13/2013		UR
Sam Industrias S. A.	Machinery and Industrial Equipment	01/07/2007	02/26/2008	B
Sansuy S.A. Indústria e Plásticos	Petrochemicals	12/20/2005		UR
Tecnosolo S. A.	Construction and Decoration Materials	08/03/2012		UR
Teka Tecelagem Kuehnrich S. A.	Textile and Apparel	10/26/2012		UR
Viação Aérea S.P. S. A. Vasp	Air Transportation	07/01/2005	11/04/2008	B
Viver Incorp. e Construt. S. A.	Construction and Decoration Materials.	06/16/2016		UR
Wetzel S. A.	Metallurgy/Steel Industry	02/03/2016		UR

Legend: UR = Under Recovery; B = Bankrupt; R = Recovered

Source: Adapted from Santos et al. (2018).

Financial data was obtained from the Economatica database (economatica.com). Financial information available at least three years before the bankruptcy filing was selected. The dataset comprised 681 companies. The database was structured in the Stata statistical software, using the reshaping procedure to organize variables and segregate data by company and year. Formulas were implemented in Stata, and the file was exported to Excel for coding.

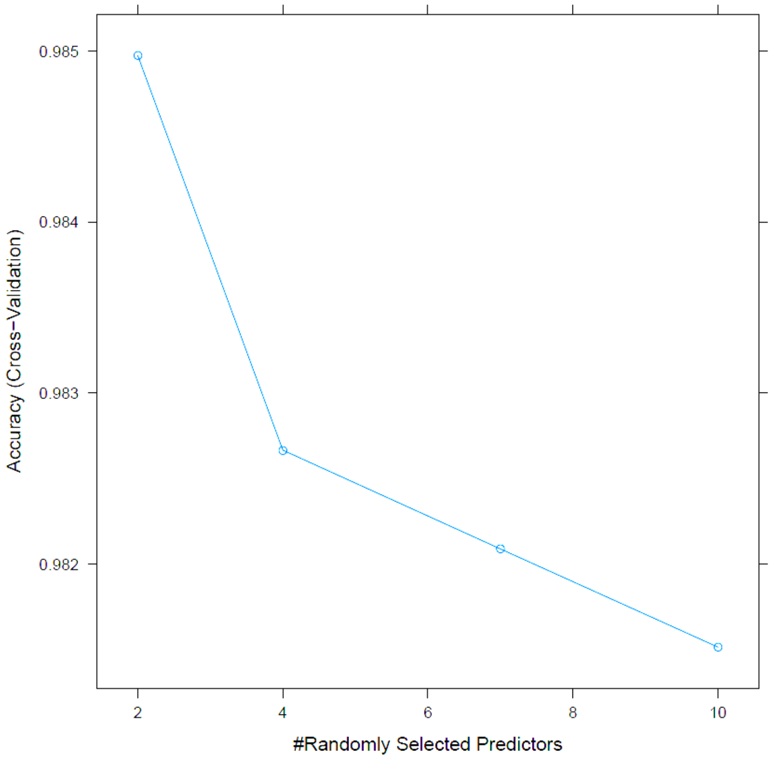
Entities with missing data were excluded from the revised database, resulting in 10,064 observations of companies/years. A distinct variable was assigned to companies that filed for bankruptcy, assigned a value of 1 for the bankruptcy year and 0 otherwise, as well as for entities that did not file for bankruptcy.

The Random Forest model was implemented on a standard MacBook Air (4GB DDR3L RAM, 64GB flash storage, Intel Core i5 1.7GHz processor, and Mac OS X operating system) equipped with R statistical software version 4.4.0 (The R Foundation, 2024). Similar to Barboza, Kimura, and Altman (2017), all variables were used in their original values without specific or special data treatment in the sample to assess the automatic application of the model. The results are presented below.

4 RESULTS AND ANALYSIS

The model's performance evaluation was conducted using 70% of the dataset for training and the remaining 30% for testing, following Barboza, Kimura, and Altman (2017). Figure 1 presents the model's accuracy capability, demonstrating a validation of the Random Forest model for application in Brazilian companies with an accuracy rate of 98%. This high level of accuracy is supported by Booth, Gerding, and McGroarty (2014); Cano et al. (2017); Kruppa et al. (2013); Yeh, Chi, and Lin (2014); Barbosa, Kimura, and Altman (2017).

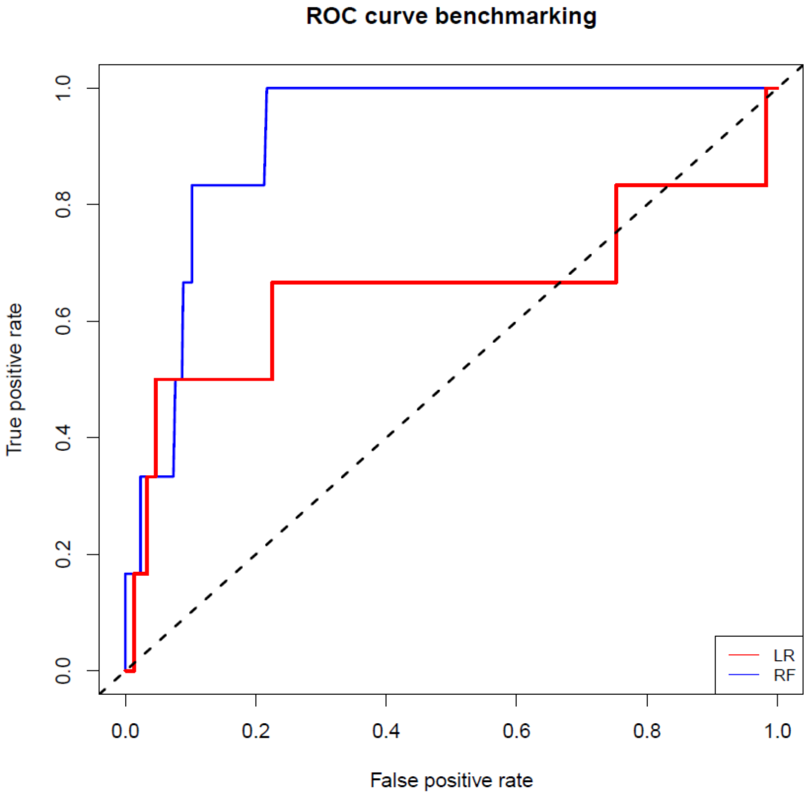
Figure 1 - Degree of Accuracy of the Random Forest Technique



Source: Research Results.

In contrast to logistic regression, which currently stands out as one of the most widely accepted and employed techniques due to its conventional statistical nature (Cano et al., 2017), the Random Forest model exhibits a significantly higher level of accuracy. This difference is evident in Figure 2, where the ROC curve presents a measure of total independent accuracy, where the value of the area below 0.5 is not valid as it is considered random, and values close to 1 demonstrate the model's ability to make accurate predictions. The comparison presented by Figure 2 clearly indicates the superior performance of the Random Forest model compared to the conventional logistic regression model.

Figure 2 - Comparison between Random Forest and Logistic Regression Techniques

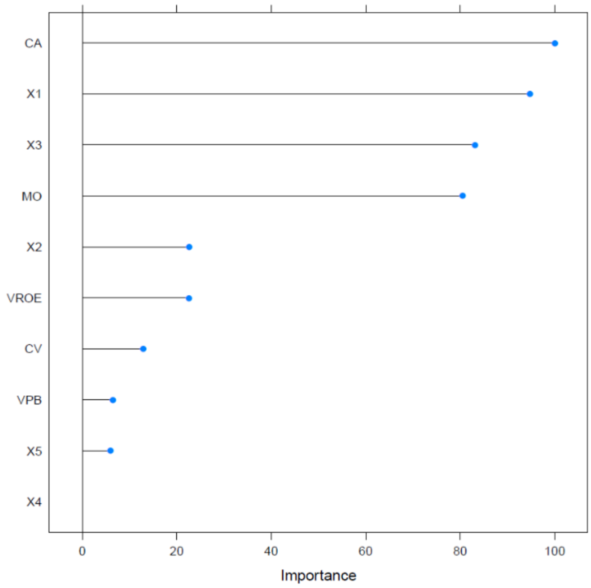


Source: Research Results.

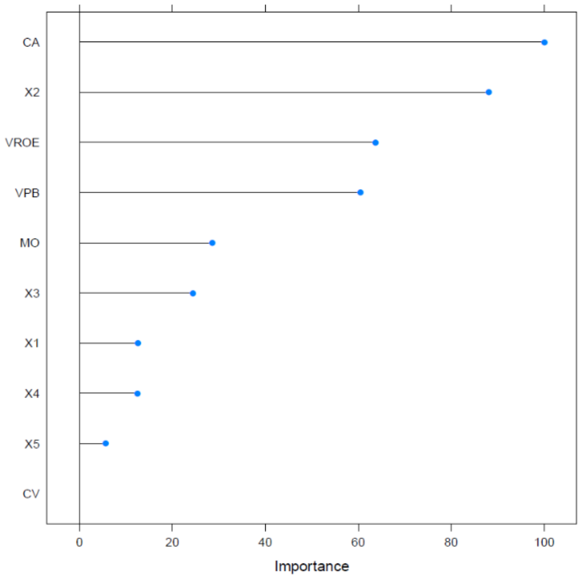
An additional crucial aspect to consider concerns the variables that exert the most significant influence on the model. Abdou (2009) emphasizes the importance of identifying variables that may indicate potential financial difficulties, allowing companies to monitor specific characteristics of their operations to mitigate and prevent financial risks. The relevant variables for the analyzed models are presented in Figures 3 and 4.

Figure 3: Influence of Variables in Random Forest

Figure 4: Importance of Variables in Logistic Regression



Source: Research Results.



Source: Research Results.

Regarding the variables that stand out the most in predicting bankruptcy, the Random Forest model identifies asset growth (CA), liquidity (X1), productivity (X3), and operating margin (MO) as the most influential factors, aligning with Altman's (1968) findings. On the other hand, variables such as leverage (X4), asset turnover (X5), and the change in price-to-book value ratio (VPB) exhibit minimal weights in the model, indicating lesser importance.

Comparing with the most influential variables in the logistic regression model, asset growth (CA) maintains its prominence as the most influential factor, followed by profitability (X2), the change in return on equity (VROE), and the change in the price-to-book value ratio, which play significant roles in bankruptcy prediction. Notably, sales growth (CV) did not show influence in logistic regression, despite being the main influencer of profit, which is directly related to asset growth. This highlights the notion that models serve as tools to support management and organizational analysis but should not be the sole criterion for evaluating a company's performance.

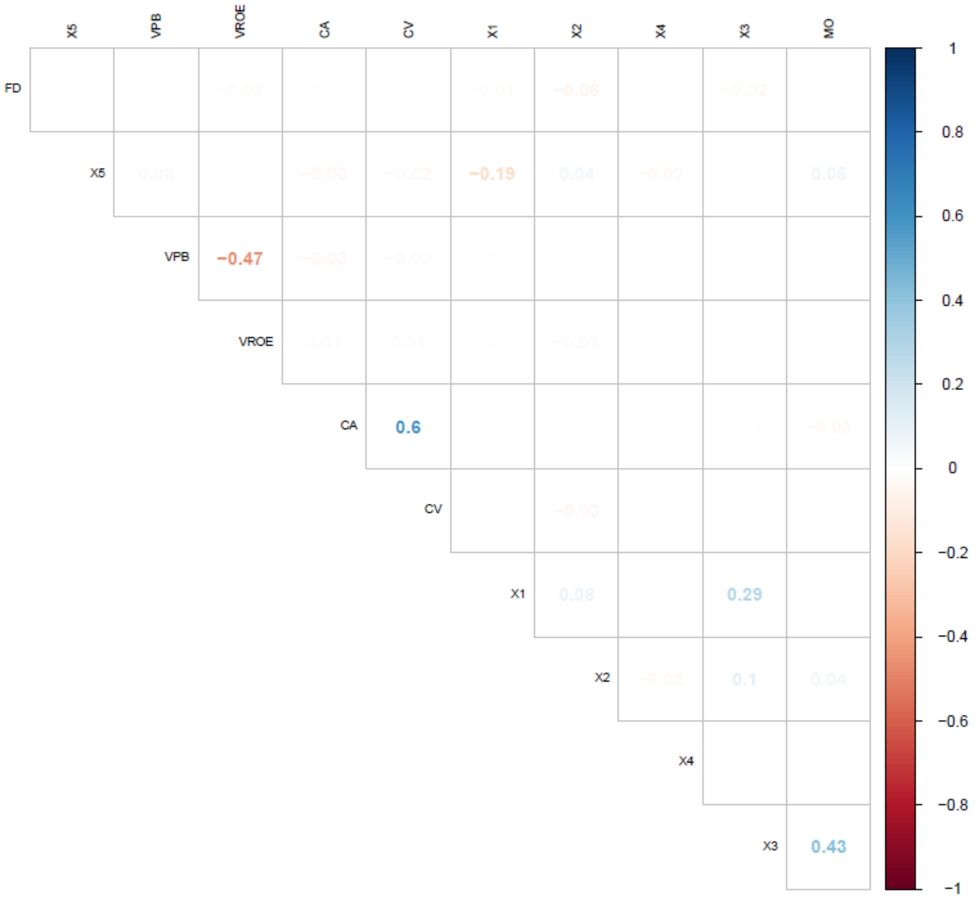
Figures 3 and 4 demonstrate a significant disparity in impact between the top four variables and the others, with the last variable showing null significance in both techniques, despite being distinct variables. This suggests a logic of categorization and exclusion in the system, which does not imply disregarding the analysis of any variable during the monitoring of an organization's financial results. Although this approach is an integral part of the technique's operation, it should not be evaluated solely based on the model's comprehensive results.

Following this logic, the correlation between the variables was evaluated, as shown in Figure 5. The results reveal that the variables in the Random Forest model, despite showing low influence, have strong correlations with other variables in the model. For example, asset turnover (X5), considered a variable of low influence, has a moderate negative correlation with liquidity (X1) (-0.19) and exhibits low positive correlations with profitability (X2) (0.04) and operating margin (MO) (0.06), the latter two recognized for their substantial influence in the model. Notably, the change in price-to-book value ratio (VPB), identified as a variable with low influence in the model, demonstrates a high negative correlation with the change in return on equity (VROE), emphasizing the need for a holistic evaluation of the model.

The results from Figure 5 also indicate a robust positive correlation between productivity (X3) and operating margin (MO). Liquidity (X1) exhibits a positive correlation with productivity (X3), while asset growth (CA) shows a low positive correlation with sales growth (CV). On the other hand, the change in return on equity (VROE) shows a substantial negative correlation with the change in price-to-book value ratio (VPB); meanwhile, liquidity (X1) has a moderate negative correlation with asset turnover (X5).

Based on these results, it can be inferred that the variables are significant in shaping the model, thus impacting the produced results, even though the Random Forest technique works with non-linear decision models. This suggests that changes in the configuration of the model variables can influence its performance, indicating the potential adaptation of the model to organizational data and the exploration of additional variables to enhance its effectiveness. This encourages further investigations employing a computational modeling approach applied to financial information and decisions.

Figure 5: Correlation of Study Variables



Source: Research Results.

5 CONCLUSIONS

This study investigated the effectiveness of the Random Forest technique applied to bankruptcy prediction in companies within an emerging economy. The results demonstrate the efficacy of the Random Forest model, which achieved a high accuracy rate of 98% in forecasting bankruptcies. Additionally, comparative analysis with conventional models, such as logistic regression, revealed the superiority of machine learning models over traditional statistical models. Furthermore, the analysis presented the most influential variables in both models, highlighting the importance of factors such as asset growth, liquidity, productivity, and operating margin for the health and longevity of companies.

Moreover, the study identified strong correlations between certain variables, indicating potential interactions and dependencies within the dataset. These findings suggest the need for a holistic approach to modeling and analyzing corporate financial data, considering not only individual variables but also their relationships and interactions.

This research contributes to the understanding of bankruptcy prediction models and underscores the importance of adopting robust statistical techniques based on machine learning for evaluating and monitoring corporations, which require enhanced decision-making processes. Future research could explore additional variables and alternative modeling approaches to further improve the accuracy and reliability of bankruptcy prediction models, as this study was limited to a single machine learning technique for analysis. Additionally, the use of robust databases is recommended, given the sample limitation of the present study. Furthermore, investigating the applicability of these models in different economic contexts and industries could provide valuable insights for both professionals and policymakers.

This study contributes by providing executives and shareholders with valuable information about the inherent risks in company operations and may prompt reflections to favor the process of informed decision-making that strengthens firms' financial health and, consequently, promotes social development. Additionally, this research contributes to the discussion of bankruptcy prediction and the use of artificial intelligence computational methodologies in the organizational financial environment.

Acknowledgements

This study was financed in part by the Fundação de Amparo à Pesquisa de Minas Gerais – Brazil (FAPEMIG) and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) – Finance code 001.

REFERENCES

ABDOU, Hussein A. **Genetic programming for credit scoring**: The case of Egyptian public sector banks. Expert systems with applications, v. 36, n. 9, p. 11402-11417, 2009. https://doi.org/10.1016/j.eswa.2009.01.076

ALTMAN, E. I. **Financial ratios, discriminant analysis and the prediction of corporate bankruptcy**. The Journal of Finance, v. 23 n. 4, p. 589–609, 1968. doi:10.1111/j.1540-6261.1968.tb00843.x

ALTMAN, E. I., MARCO, G.; VARETTO, F. **Corporate distress diagnosis**: Comparisons using linear discriminant analysis and neural networks (the Italian experience). Journal of Banking & Finance, v.18, n.3, p. 505–529, 1994. doi:10.1016/0378-4266(94)90007-8

B3 – BRASIL, BOLSA, BALCÃO. **Empresas com listagem cancelada no mercado de bolsa.** Disponível em: <http://www.b3.com.br/pt_br/produtos-e-servicos/negociacao/renda-variavel/acoes/consultas/empresas-com-listagem-cancelada-no-mercado-de-bolsa/>.

BARBOZA, F., KIMURA, H., ALTMAN, E. **Machine learning models and bankruptcy prediction**. Expert Systems with Applications, v.83, p. 405–417, 2017. doi:10.1016/j.eswa.2017.04.006

BOOTH, A., GERDING, E., MCGROARTY, F. **Automated trading with performance weighted random forests and seasonality.** Expert Systems with Applications, v.41, n.8, p. 3651–3661,2014.doi:10.1016/j.eswa.2013.12.009

BREIMAN, L. **Random forests**. Machine Learning, v. 45, n.1, p 5–32, 2001. doi:10.1023/a:1010933404324

CALDERONI, L., FERRARA, M., FRANCO, A., AIO, D. **Indoor localization in a hospital environment using Random Forest classifiers**. Expert Systems with Applications, v. 42, n.1, p. 125–134, 2015. doi:10.1016/j.eswa.2014.07.042

CANO, G., GARCIA-RODRIGUEZ, J., GARCIA-GARCIA, A., PEREZ-SANCHEZ, H., BENEDIKTSSON, J. A., THAPA, A., BARR, A. **Automatic selection of molecular descriptors using random forest:** Application to drug discovery. Expert Systems with Applications, v.72, p.151–159, 2017.doi:10.1016/j.eswa.2016.12.008

CARTON, R., HOFER, C. **Measuring organizational performance.** Edward Elgar Publishing, 2006.

CASEY, C., BARTCZAK, N. **Using Operating Cash Flow Data to Predict Financial Distress**: Some Extensions. Journal of Accounting Research, v.23, n.1, p.384, 1985. doi:10.2307/2490926

CHIAVEGATTO FILHO, A. D. P. **Uso de big data em saúde no Brasil:** perspectivas para um futuro próximo. Epidemiol. Serv. Saúde, Brasília, v.24, n.2: p. 325-332, abr-jun 2015.

DAVIDOV, D.; TSUR, O.; RAPPOPORT, A. **Enhanced Sentiment Learning Using Twitter Hashtags and Smileys.** Coling 2010: Poster Volume, p. 241–249, Beijing, August 2010

GUIMARÃES E MOREIRA. **Previsão de insolvência**: Um modelo baseado em índices contábeis Com utilização da análise discriminante*. R. Econ. contemp., Rio de Janeiro, v. 12, n. 1, p. 151-178, jan./abr. 2008

HEO, J., YANG, J. Y. **AdaBoost based bankruptcy forecasting of Korean construction companies**. Applied Soft Computing, v.24, p. 494–499. 2014. doi:10.1016/j.asoc.2014.08.009

KIM, S. Y., UPNEJA, A. **Predicting restaurant financial distress using decision tree and AdaBoosted decision tree models**. Economic Modelling, v.36, p. 354–362, 2014. doi:10.1016/j.econmod.2013.10.005

KOHAVI, RON PROVOST, FOSTER. **Glossary of terms**. Machine Learning, v.30, n.2-3:, p. 271–274, 1998.

KRUPPA, J., SCHWARZ, A., ARMINGER, G., ZIEGLER, A. **Consumer credit risk: Individual probability estimates using machine learning.** Expert Systems with Applications, v.40, n.13, p. 5125–5131, 2013. doi:10.1016/j.eswa.2013.03.019

LIANG, D., LU, C.-C., TSAI, C.-F., SHIH, G.-A. **Financial ratios and corporate governance indicators in bankruptcy prediction:** A comprehensive study. European Journal of Operational Research, v.252, n.2, p.561–572, 2016. doi:10.1016/j.ejor.2016.01.012

MAIONE, C., DE PAULA, E. S., GALLIMBERTI, M., BATISTA, B. L., CAMPIGLIA, A. D., JR, F. B., BARBOSA, R. M. **Comparative study of data mining techniques for the authentication of organic grape juice based on ICP-MS analysis.** Expert Systems with Applications, v.49, p. 60–73, 2016. doi:10.1016/j.eswa.2015.11.024

PRATES. P. R. **Do Bastão de Esculápio ao Caduceu de Mercúrio**. Arq. Bras. Cardiol. v.79, n.4, São Paulo: 2002. http://dx.doi.org/10.1590/S0066-782X2002001300014

SAMUEL, A. L. **Some Studies in Machine Learning Using the Game of Checkers.** IBM Journal of Research and Development, v. 3, n.3, p. 210–229, 1959. doi:10.1147/rd.33.0210

SANTOS, V. S.; MÁRIO, P. C.; AGUILAR, D. Z.; JUPETIPE, F. K. N. **Assertividade dos Modelos de Previsão de Insolvência Aplicados a Companhias de Capital Aberto em Recuperação Judicial**. EGEN - Encontro de Gestão e Negócios. Uberlândia-MG, 2018.

SHOKRI, R., STRONATI, M., SONG, C., SHMATIKOV, V. **Membership Inference Attacks Against Machine Learning Models.** 2017 IEEE Symposium on Security and Privacy (SP), 2017. doi:10.1109/sp.2017.41

SIGILIANO, D.; FAUSTINO, E. **NETFLIX**: Sistemas de Recomendação Inteligentes. Revista Tecer - Belo Horizonte – vol. 9, nº 16, maio de 2016.

TASINAFFO, P. M. **Um breve histórico do desenvolvimento da lógica matemática e o surgimento da teoria da computação.** Anais do 14O Encontro de Iniciação Científica e Pós-Graduação do ITA – XIV ENCITA / 2008 Instituto Tecnológico de Aeronáutica, São José dos Campos, SP, Brasil, Outubro, 20 a 23, 2008.

TEIXEIRA, J. **Metáforas da crise cotidiana: os media e a veiculação da crise grega in Revista Investigações** – Linguística. Edição Temática - 35 anos de Metáforas da Vida Cotidiana. v. 28, n. 2, julho/2015. Universidade Federal de Pernambuco. ISSN Edição Digital 2175-294X

THE R FOUDATION. **R**: the R project for statistical computing. Vienna: The R Foundation, 2024.

TSAI, C.-F., HSU, Y.-F., YEN, D. C. **A comparative study of classifier ensembles for bankruptcy prediction.** Applied Soft Computing, v. 24, p. 977–984. 2014. doi:10.1016/j.asoc.2014.08.047

TURING, A. M. **Computing machinery and intelligence.** Mind, v. LIX, n.236, p.433–460, 1950. doi:10.1093/mind/lix.236.433

WANG, G., MA, J., YANG, S. **An improved boosting based on feature selection for corporate bankruptcy prediction.** Expert Systems with Applications, v.41, n.5, p. 2353–2361, 2014. doi:10.1016/j.eswa.2013.09.033

XIAO, H., XIAO, Z., WANG, Y. **Ensemble classification based on supervised clustering for credit scoring.** Applied Soft Computing, v.43, p. 73–86. 2016. doi:10.1016/j.asoc.2016.02.022

YEH, C.-C., CHI, D.-J., LIN, Y.-R. **Going-concern prediction using hybrid random forests and rough set approach.** Information Sciences, v. 254, p. 98–110, 2014. doi:10.1016/j.ins.2013.07.011