05, 06, 07 e 08 de novembro de 2024
ISSN 2177-3866

Evaluation of Machine Learning Models in Predicting Bankruptcies in Brazilian Companies

JOSÉ ERASMO SILVA
UNIVERSIDADE FEDERAL DA BAHIA (UFBA)

LUIS PAULO GUIMARÃES DOS SANTOS
UNIVERSIDADE FEDERAL DA BAHIA (UFBA)

SHEIZI CALHEIRA DE FREITAS
UNIVERSIDADE FEDERAL DA BAHIA (UFBA)

CÉSAR VALENTIM DE OLIVEIRA CARVALHO JUNIOR
UNIVERSIDADE FEDERAL DA BAHIA (UFBA)

Evaluation of Machine Learning Models in the Prediction of Bankruptcies in Brazilian Companies

Introduction

Corporate bankruptcy forecasting is vital for various stakeholders, from managers and investors to employees and customers. These forecasts help mitigate financial risks and play a crucial role in maintaining global economic stability. Historically, the prediction of bankruptcy has been intensively explored since the 1960s, reflecting its importance in the economic development of nations (Ferren & Kurniadi, 2023; Yotsawat, Phodong & Wattuya, 2023).

A key component in predicting bankruptcy is assessing the financial health of companies, which is relevant to various stakeholders, including investors, regulators, and other companies that may be affected by their operations. Predicting bankruptcy and financial vulnerability becomes essential, as these conditions can precede bankruptcies and liquidations. This is particularly relevant for publicly traded companies listed on stock exchanges, where a company's financial health can have significant implications for the market as a whole (El Madou, Marso, El Kharrim & El Merouani, 2023; Letkovský, Jencová & Vasanicová., 2024; Muslim & Dasril, 2021; Platt & Platt, 2002).

However, researchers emphasize that diverse economic environments require adaptive forecasting models due to their unique characteristics. This necessity is particularly pronounced in sectors such as construction, which have distinct financial and operational risks and are deeply influenced by economic and regulatory variables (Giriūniene, Girunas, Morkunas & Brucaite, 2019; Thanh-Long, Tran-Minh & Hong-Chuong et al., 2022).

In addition, regional studies, such as those conducted by Visegrad countries, reveal that financial variables such as profitability and liquidity are critical, but models vary significantly between countries. This variation suggests developing more flexible models adapted to specific economic and business contexts (Kovacova, Kliestik, Valaskova, Durana & Juhaszova, 2019).

Despite advances in machine learning and deep learning tools, which provide high accuracy rates in predicting bankruptcies, as demonstrated, there is a significant gap in the application of these models in less studied environments, such as Africa, South America, and Southeast Asia (Máté, Raza & Ahmad., 2023).

Considering the specificity of the Brazilian market and the importance of developing bankruptcy prediction models that are accurate and adjusted to local characteristics, this study is justified by the need to adapt and refine such models for the context of B3, the Brazilian stock exchange. The objective of this research is to explore the applicability of six machine learning models - including Random Forest, XGBoost, LightGBM, CatBoost, Support Vector Machine (SVM), and Logistic Regression - exclusively in the Brazilian market, paying particular attention to the adaptation of these models to different industrial sectors listed on B3. This study is expected to contribute to the financial stability of Brazilian companies, minimizing the economic and social impact of bankruptcies in the country and providing relevant information for investors, regulators, and policymakers.

Literature review

Fluctuations in the stock market are seen as indicators of global economic health, reflecting systemic economic conditions. These fluctuations capture immediate responses to economic and political uncertainties, providing insights into financial stability and global economic challenges. (Khalil & Burn, 2023).

Many experts argue that bankruptcy prediction is transitioning from traditional techniques like discriminant analysis and logistic regression to modern methods such as

Ensemble models, Neural Networks, SVM, RNN, and CNN. This shift reflects the evolution towards machine learning and deep learning, which offer more intricate and precise analyses by capturing nonlinear relationships and complex data interactions beyond the capabilities of traditional models (Qu, Quan, Lei, & Shi, 2019).

A study by Kristanti and Dhaniswara (2023) analyzed industrial companies on the Indonesian stock exchange from 2017 to 2021. The study compared Logit and Artificial Neural Network (ANN) models for predicting financial events. Surprisingly, the traditional Logit model outperformed the modern ANN model, achieving 98% accuracy, 94.20% sensitivity, and 99.30% specificity, while the ANN model scored 82.50% accuracy, 84% sensitivity, and 82% specificity.

El Madou et al. (2023) reviewed bankruptcy prediction methodologies from 1966 to 2022, categorizing them into Simple Models (traditional statistical approaches), Hybrid Models (combining statistical methods), and Set Methods (integrating multiple predictions). Their study highlights a shift from simple to complex models, reflecting efforts in financial research to enhance forecasting accuracy and early bankruptcy risk identification.

Wang, Kräussl, Zurad, and Brorsson (2023) in a study examining the impact of feature engineering on bankruptcy predictions using data from the Luxembourg Commercial Registers (2011-2021), the researchers employed the LightGBM algorithm to select significant traits. This approach led to substantial improvements in the model's Area Under the Curve (AUC), underscoring the crucial role of careful trait selection and the value of feature engineering in enhancing bankruptcy prediction accuracy.

In a study conducted by Syafei and Efrilianda (2023), the effectiveness of machine learning techniques in improving the prediction of bankruptcies was explored. Using a sample of 6,819 Taiwanese companies collected between 1999 and 2009, the authors initially used the XGBoost algorithm to select the characteristics. Then, they applied LightGBM to classify the companies, considering their condition of bankruptcy or non-bankruptcy. Due to the inherent imbalance in the failure samples, a Random oversampling procedure was implemented to balance the classes, a common practice in this type of analysis. This adjustment improved the model's accuracy from 96.583% to 98.916% after balancing, with cross-validation at ten k-folds. This study demonstrates the applicability of machine learning algorithms in the early detection of bankruptcies and underscores the importance of data balancing for model accuracy in unbalanced data scenarios.

Liashenko, Kravets, and Kostovetskyi (2023) addressed the persistent challenge of data imbalance in predicting U.S. business failures by using a time series covering the period from 1980 to 2014. The authors employed several machine learning models, including Random Forest, which had one of the highest accuracy, precisely 91.61%. Notably, the ANN and the Decision Trees demonstrated excellence in Area Under the Curve (AUC), with 78.33% and 79.79% results. They highlighted that balancing techniques such as near-miss and random undersampling are vital to mitigate the risk of overfitting in models relative to the majority class, a prevalent condition in predicting bankruptcies due to their lower comparative frequency.

Muslim and Dasril (2021) explored the prediction of bankruptcies among Polish firms from 2000 to 2012, using a stacking approach to refine the forecasts. The initial selection of features was made with XGBoost, focusing on high-relevance attributes determined by a filter with a weight value of 10. The stacking model, composed of K-Nearest Neighbors (KNN), Decision Tree, SVM, and Random Forest with LightGBM as a meta-learner, outperformed the base models, achieving an accuracy of 97%. This study illustrates the effectiveness of ensemble learning methods in overcoming the limitations of individual models, offering robust results for financial decision-making.

Thanh-Long et al. (2022) evaluated bankruptcy prediction accuracy in the construction sector using 1,262 observations from 44 bankrupt and 1,218 solvent companies. They implemented two neural network configurations: the Retropropagation Neural Network (BPNN) and SMOTE-BPNN. The SMOTE-BPNN model improved prediction accuracy from 79.9% to 84.1% and effectively addressed class imbalance and selection biases, enhancing bankruptcy risk modeling in the construction sector.

Thanh-Long et al. (2022) emphasized the uniqueness of the construction industry, noting that it is distinguished from other industries due to its exposure to a complex set of financial and operational risks. The study details how intrinsic uncertainty and dangers arising from technical, human, and natural variables define the sector's risk profile. Furthermore, the authors note that the financial health of construction companies is susceptible to the prevailing economic environment, government regulations, and fluctuations in the business cycle. This sensitivity is attributed to the industry's cyclical nature, which frequently faces periods of boom and bust that align with the overall economic dynamics. The discussion is crucial to understanding how these factors directly impact the solvency of companies in the sector, offering vital information for risk management and strategic planning in the context of bankruptcy prediction.

In the study conducted by Thilakarathna, Dawson, and Edirisinghe (2022), the prediction of business bankruptcy was investigated through a sophisticated approach in feature engineering. Using an extensive dataset encompassing more than 14,000 companies from the UK and Ireland, which was analyzed over ten years, the study explored 29 financial indicators as predictors. The authors developed and implemented a Deep Neural Network (DNN) model to improve the accuracy and anticipation of failure predictions. The study's remarkable results highlight the model's ability to achieve an AUC score of 0.897 when predictions are made one year before eventual bankruptcy. This performance highlights the model's effectiveness compared to previous methodologies and solidifies the importance of advanced machine learning techniques in financial predictive analytics.

In a study conducted by Sayed and Khalil in 2022, the significant impact of cash reserves on bankruptcy risk in corporate contexts was explored. The study focused on companies listed on the Egyptian Stock Exchange (EGX) from 2015 to 2019. The researchers used a sample of 340 observations from 68 companies. They applied econometric analysis methods, including panel-corrected standard errors (PCSE) and feasible generalized least squares (FGLS), to assess the nuances of this relationship. Additionally, the study looked into the moderating role of corporate social responsibility (CSR) practices in this dynamic. The findings revealed that cash reserves are crucial in mitigating bankruptcy risk and that implementing CSR practices significantly strengthens this influence. This finding not only underlines the importance of prudent financial management but also highlights how integrating ethical and responsible practices can amplify businesses' financial stability. Therefore, the study offers significant contributions to the literature on corporate finance, suggesting that social responsibility and liquidity management are essential for organizational resilience in emerging markets.

In the study conducted by Sulistiani, Widodo, and Nugraheni (2022), corporate bankruptcy analysis was critical in financial decision-making, focusing on state-owned enterprises in Indonesia between 2010 and 2019. Using annual financial reports, the study applied sophisticated machine learning methodologies, including SVM and ANN, to predict potential bankruptcies. The authors implemented the Principal Component Analysis (PCA) Dimensionality Reduction technique to address technical challenges such as overfitting and efficient feature selection. Additionally, Synthetic Border Minority Oversampling (BSM) was employed to balance the data set, thus addressing the problem of imbalance between failed and non-bankrupt firms. The results showed that the combination of ANN-BSM-PCA, with 16 principal components, was particularly effective, achieving remarkable precision, accuracy, and sensitivity of 0.96, 0.934, and 0.988, respectively. This model outperformed the SVM, demonstrating the effectiveness of integrated machine learning techniques in predicting bankruptcies with high accuracy, offering a foundation for policymakers and financial managers.

In the study conducted by Premalatha, Priyanka, and Chaitya (2023), the focus was on identifying the key attributes contributing to business failure. Employing data compiled by the Taiwan Economic Journal from 1999 to 2009, which includes 6,819 companies and 96 distinct variables, the researchers utilized a novel combination of advanced analytical methods, such as variance threshold, mutual information, and Pearson correlation, along with the SMOTE technique to balance the data set. This multifaceted approach was complemented by using four machine learning classifiers: Random Forest, Decision Tree, AdaBoost, and XGBoost, each tested to assess their effectiveness in predicting bankruptcies. The results showed that the proposed strategy surpasses traditional approaches, with accuracy rates ranging between 83.34% and 97.07%, depending on the methods of attribute selection and the classifiers employed. This study highlights the relevance of sophisticated attribute selection techniques in predicting bankruptcies. It emphasizes the importance of adaptive machine-learning strategies that significantly refine predictive capabilities in complex economic scenarios.

Shah, Rao, Mehta, and Kurhade (2022) it investigated the prediction of corporate bankruptcy using a focus that highlights its critical importance to financial institutions and its broader impact on the economy. Using the Random Forest model, the study explored an extensive dataset of 10,000 records and 64 attributes extracted from the Polish bankruptcy register to identify companies at risk of bankruptcy. The approach adopted achieved an accuracy of 97.35% in classifying companies as bankrupt or solvent. In addition, the authors employed a linear regression model to analyze and determine the optimal values of financial attributes indicative of financial stability. The study's results confirm the Random Forest model's effectiveness in accurately predicting bankruptcies and provide insights into enhancing specific financial characteristics to mitigate bankruptcy risk. This research highlights the relevance of sophisticated and adaptive modeling in predicting financial risks, offering contributions to decision-making in volatile financial environments.

The study by Lombardo et al. (2022) investigated predicting corporate bankruptcies in the American stock market. The sample comprised accounting data from 8,262 companies collected between 1999 and 2018. Machine learning models, including decision trees, random forests, gradient boosting, and neural networks, were used for the methodology. Two distinct bankruptcy forecasting tasks are explored: the first aims to predict the company's status in the coming year based on time series of accounting data, while the second focuses on estimating probabilities of survival over several years. The results are critically discussed, highlighting metrics such as AUC, precision, sensitivity, and type I and II errors as proposed benchmarks for future studies.

In the study of Bittencourt and Albuquerque (2020) companies' bankruptcy was analyzed with a focus on feature engineering. The main objective was to explore the Causal Forests methodology to identify variables that could influence company bankruptcy. The sample used included quarterly and sectoral accounting data from 1,247 companies, 66 bankrupt. The research highlights the importance of considering heterogeneity between companies and questions the effectiveness of traditional models such as discriminant analysis and logit. The Causal Forests methodology is presented as a solution to these challenges, allowing flexible modeling with high levels of interactions and dimensions.

In the study by Lott, Tenenwurcel, and Camargos (2021), Brazilian companies' capital structure on B3 was examined with a focus on determinants of indebtedness and insolvency risk. The sample consisted of 233 companies, with financial data collected between 2011 and 2016. Altman's Z2 failure prediction index and a multiple regression model estimated by ordinary least squares (OLS) are used. The results indicate significant differences in the determinants of indebtedness between companies with and without risk of bankruptcy.

Companies at risk of bankruptcy have a positive relationship between long-term and total debt and profitability and risk. For healthy companies, the relationship is negative with profitability and positive with risk and size. The study has limitations regarding the generalizability of the results and contributes by providing new empirical evidence on a controversial topic in financial theory.

The study by Rodríguez-Masero and López-Manjón (2020) examined the usefulness of operating cash flow (FCO) for predicting bankruptcy in medium-sized companies. The research employed logit analysis to build a data-driven model of the financial statements. The sample was composed of medium-sized Spanish companies included in SABI. The study identified an integrated function for several ratios, including information from the cash flow statement, helpful in assessing whether a company can be classified as solvent or insolvent in the future. The results suggest that the model can predict the probability of a medium-sized company facing financial problems three years before they occur, allowing the implementation of corrective measures. The study contributes to the literature by focusing on medium-sized companies and using operating cash flow as a significant variable for predicting bankruptcy.

In the study conducted by Prado, Carvalho, Benedicto, and Lima(2019), the Credit Risk Faced by Public Enterprises in Brazil was analyzed to identify the economic and financial indicators that best contribute to the accuracy of credit granting analyses and evaluate the most accurate techniques for predicting business bankruptcy. The sample comprised 121 companies from various sectors, 70 solvents, and 51 insolvents. Three methods were used to predict insolvency: Discriminant Analysis, Logistic Regression, and Neural Networks. The most relevant economic and financial indicators identified were the need for working capital over net income, liquidity thermometer, return on equity, profit margin, debt ratio, and equity over assets. The neural network model showed the highest accuracy, corroborated by the ROC curve.

Heo and Yang (2014) used the AdaBoost algorithm to predict the bankruptcy of Korean companies in the construction sector. According to the authors, this sector has specific characteristics that differ from others, so traditional bankruptcy prediction models are ineffective. In addition to AdaBoost, the authors tested the ANN, SVM, Decision Tree, and Z-Score models, with the best results found for AdaBoost and the worst for Z-Score.

Yu, Miche, Séverin, and Lendasse (2014) in addition to traditional approaches, machine learning methods, such as Decision Trees, fuzzy theory, genetic algorithms, and SVM, have been applied to bankruptcy prediction since the 1990s. However, all of these methods, including traditional financial models, suffer from problems of strict hypotheses, poor generalizability, poor forecasting accuracy, low learning rate, and slow computational time. To overcome these limitations, the authors propose using the Extreme Learning Machine (ELM), a machine learning approach known for its efficiency and ability to train models quickly. The authors considered ELM a potential solution to the problems above, suggesting that it may offer improvements in hypotheses, generalization, accuracy, learning rate, and computational time. The maximum accuracy found achieved by the authors was above 90% for most tests.

Shetty, Musa, and Brédart (2022) used Neural Networks, SVM, and XGBoost to predict bankruptcy in small Belgian firms. Despite using only three explanatory variables, Return on Assets, Current Ratio, and Solvency Ratio, the forecasts were accurate by 83%.

Chen, Chen, and Shi (2020) proposed two new methods, Bagged-pSVM and Boosted-pSVM, for bankruptcy prediction, using companies from different countries as samples. The authors highlighted that for studies of this type, there is great difficulty in obtaining data labeled at the instance level. To overcome this, the authors propose to approach the prediction of bankruptcies from the perspective of learning with label ratios (LLP), in which unlabeled training data is provided in different "grants," and only the proportion of instances belonging to a specific class is known. The methods were promising and achieved accuracy above 82%. The authors did not provide much information about the variables used; they were only different in different countries, with a minimum of 14 in Australia and a maximum of 64 in Poland.

Baldissera, Fiirst, Rovaris, and Dall'Asta (2020) state that understanding companies' characteristics before filing for bankruptcy or judicial reorganization is vital to providing accurate information to stakeholders. They analyzed the capital structure of the companies on B3 in the five years before these requests. They used a sample of 36 companies, half solvent and half insolvent. It was found that before these orders, the capital structure followed the pecking order theory. These companies have insufficient assets for their obligations, resorting to external financing since they cannot raise their resources.

Methodology

Sample

All companies in the Comdinheiro platform's database were initially surveyed, with data available between 01/01/2000 and 12/31/2023. After the survey, companies belonging to the "financial" and "financial and other sectors" were excluded. Some companies present data for the entire period; however, some newer companies have shorter periods. All of them were kept in the sample, totaling 564 companies with 23385 observations. Table 1 shows the number of companies for each registration situation with the CVM.

Table 1

Number of companies per situation at the CVM

Situation	Quantity	Percentage (%)
Cancelled	112	19.86
In receivership	1	0.18
In judicial reorganization	2	0.35
In judicial reorganization or equivalent	18	3.19
Broke	6	1.06
Operational phase	414	73.40
Pre-operational phase	5	0.89
Extrajudicial liquidation	3	0.53
Paralyzed	3	0.53
Total	564	100

Source: CVM (2023) and Comdinheiro (2023)

It should be noted that bankruptcy is relatively rare on the stock exchange. This is because other measures can be taken before reaching this outcome, such as judicial reorganization or acquisition by another company. In this way, even without culminating in bankruptcy, the company can, in the previous phases, generate losses for the interested parties. In the present study, the following categories were adopted as criteria for the classification of "bankruptcy" "in judicial reorganization or equivalent," "bankrupt," "extrajudicial liquidation," "in extrajudicial reorganization," and "in judicial liquidation," totaling 27 companies (Guimarães & Resende Filho, 2018).

It is worth noting that some studies have different perspectives on insolvency, such as considering the company's Uncovered Liabilities. Thus, a company can be classified as insolvent if it has Negative Shareholders' Equity at the end of the fiscal year (Guimarães & Alves, 2009). Table 2 provides an overview of the distribution of these companies by industry.

Table 2

Number of companies by sector

Sector	Quantity	Percentage (%)
Industrial Goods	73	12.94
Communications	9	1.60
Construction and Transportation	17	3.01
Cyclical Consumption	119	21.10
Non-cyclical consumption	50	8.87
Hotels & Restaurants	1	0.18
Basic Materials	51	9.04
Not Rated	89	15.78
Other	20	3.55
Oil, Gas & Biofuels	14	2.48
Chemists	1	0.18
Health	26	4.61
Information Technology	22	3.90
Telecommunications	9	1.60
Public Utility	63	11.17
Total	564	100

The Cyclical Consumption sector has the most companies, representing approximately 21.10%. In contrast, the Hotels & Restaurants and Chemicals sectors have the fewest companies, each with only one company in the sample.

Models

According to Pereira and Martins (2016), the predominant methodologies since 1968 have been the statistical models of Linear Discriminant Analysis. Logit and Probit models also had some applications but to a lesser extent. From the end of the 1980s, machine learning models emerged, emphasizing neural networks, which have become increasingly relevant. They also point out that one of the limitations of the studies before 1980 is that they were based only on accounting data recorded at historical cost. However, from the 1980s onwards, they began to use cash flow-based metrics. Table 3 briefly surveys the methodologies used to predict bankruptcy and insolvency.

Table 3

Summary of the methodologies used in the studies surveyed

Author(s)	Methodology Used
Premalatha et al. (2023)	They identified essential attributes for failure using a combination of variance threshold, mutual information, Pearson correlation, and the SMOTE technique. They used Random Forest, Decision Tree, AdaBoost, and XGBoost as classifiers.
Shah et al. (2022)	The focus was on predicting corporate bankruptcy using the Random Forest model based on a Polish bankruptcy dataset, which was accurate at 97.35%. They also included a linear regression model to determine the optimal value of financial attributes.
Bittencourt and Albuquerque (2020)	Analysis of business bankruptcy using the Causal Forests methodology to identify influencing variables.

Liashenko et al. (2023)	She used data balancing methods to address the imbalance problem in training data for predicting bankruptcies.
Muslim and Dasril (2021)	Study of Polish companies using the importance of XGBoost for feature selection and a stacking model composed of KNN, decision tree, SVM, and Random Forest, with the meta-learner being LightGBM.
El Madou et al. (2023)	Discuss approaches and methods to predict the probability of bankruptcy, categorizing them into Simple Models, Hybrid Models, and Set Methods.
Máté et al. (2023)	Use AdaBoost algorithms, decision trees, gradient boosting, logistic regressions, naïve Bayes, random forests, and support vector machines in Pakistani companies.
Kovacova et al. (2019)	Analysis of more than 100 bankruptcy prediction models in the Visegrad group (V4) countries, identifying relevant variables such as profitability ratios, current ratios, and liabilities-to-total assets ratio.
Kristanti and Dhaniswara (2023)	Analysis of companies in the industrial sector on the Indonesian stock exchange using a Logit model, highlighting the superiority of this model over others in predicting bankruptcy.
Baldissera et al. (2020)	Study companies' capital structure on B3 before a bankruptcy or judicial reorganization filing, using theories such as the *trade-off* and *the pecking order.*
Barboza et al. (2017)	It tests eight techniques for predicting bankruptcy in more than 13,000 U.S. companies, including Altman's Z-Score, logistic regression, and decision tree-based models (Boosting and Bagging).
Lombardo et al. (2022)	Predicting bankruptcies using unlabeled data and learning from label ratios (LLP)
Qu et al. (2019)	Discuss the transition from old to new techniques in predicting failure, including Ensemble models, Neural Networks, SVM, RNN, and CNN.
Wang et al. (2023)	Engineering features in financial records for bankruptcy prediction, using the LightGBM algorithm to select essential features, demonstrating effectiveness in increasing AUC compared to the direct approach to economic indicators.
Syafei and Efrilianda (2023)	XGBoost was used to select the most essential characteristics, and LightGBM was used to classify bankrupt companies, using a sample of companies from Taiwan. The data imbalance was adjusted with the random oversampling procedure.
Lott et al. (2021)	Examination of the capital structure of Brazilian companies listed on B3, focusing on determinants of indebtedness and insolvency risk, using a sample of 233 companies with financial data.
Heo and Yang (2014)	Use of the AdaBoost algorithm to predict the bankruptcy of Korean companies in the construction sector. They also tested ANN, SVM, Decision Tree, and Z-Score, with AdaBoost showing the best results.
Yu et al. (2014)	They proposed using the Extreme Learning Machine (ELM) to overcome the limitations of traditional AI methods, such as decision trees and support vector machines, highlighting the efficiency of ELM in training models quickly.
Shetty et al., (2022)	Using three explanatory variables, they employed Neural Networks, SVM, and XGBoost to predict bankruptcies in small Belgian firms.
Chen et al. (2020)	They proposed the Bagged-pSVM and Boosted-pSVM methods for predicting bankruptcy and approached predicting bankruptcies from the point of view of learning with label ratios (LLP).

Table 3 shows that the evolution of the methodologies used in predicting bankruptcies reflects a significant transition from traditional statistical techniques to approaches based on artificial intelligence. Initially, linear discriminant analysis, Logit, and Probit models dominated the field. However, from the late 1980s to the 1990s, neural networks emerged as preferred methods, increasing complexity and modeling capacity. More recently, there has been a marked preference for ensemble algorithms such as Random Forest, AdaBoost, and XGBoost, highlighting a continuous search for accuracy and robustness in predictions. This pattern suggests a growing appreciation of the ability of models to deal with the complexity and heterogeneity of financial data.

Despite methodological advances, there are significant gaps in bankruptcy prediction research. First, the lack of standardization in comparisons between different methods makes it difficult to identify the most effective in various contexts. In addition, the lack of longitudinal studies that follow companies over time to validate predictions represents a missed opportunity to assess the practical applicability of models. Another critical gap is the limited exploration of the impact of regional economic and cultural variables on the effectiveness of predictive models, suggesting the need for a more globalized and contextualized approach in future research.

The analysis of the methodologies reveals interdisciplinarity and technological innovation as central aspects of bankruptcy prediction research. The convergence of knowledge from different fields, such as finance, statistics, and computer science, highlights the complexity of the challenge and the importance of collaborative approaches. The rapid adoption of advanced artificial intelligence techniques reflects the impact of technological innovation and points to a constant search for more accurate and applicable models. These advances, along with attention to data imbalance and the diversity of economic contexts, highlight a commitment to continuous improvement and the practical relevance of bankruptcy prediction research.

Given this, six machine-learning models were selected and inserted into four groups. Random Forest is a bagging algorithm that builds multiple decision trees and aggregates their results (Breiman, 2001). XGBoost is a boosting algorithm that builds decision trees sequentially, where each subsequent tree is constructed to correct the mistakes made by the previous trees (Chen et al., 2018; Chen & Guestrin, 2016). LightGBM is also a boosting algorithm that uses the gradient-based boosting strategy (Ke et al., 2017). CatBoost is another boosting algorithm that handles categorical features (Prokhorenkova et al., 2017). SVM is a margin-based algorithm that attempts to find the hyperplane that maximizes the margin between classes (Ben-Hur & Weston, 2010). Finally, logistic regression is a statistical model that uses a logistic function to model a binary dependent variable (Vapnik, 2000).

Variables

This study uses a series of variables identified from previous studies. These variables, detailed in Table 4, were selected for their relevance and frequency of use in research.

Table 4

Variables used in other studies

Variable	Description	Author(s)
READ	Liquidez Imediata	General Studies
LC	Current Liquidez	General Studies
LS	Dry liquidity	General Studies
LG	General Liquidity	General Studies
Current Assets	Assets that can be converted to cash within a year	General Studies
Current liabilities	Financial obligations due within one year	General Studies
Total Assets	Total sum of a company's assets	General Studies
Net Revenue	Total revenue minus returns, rebates, and taxes	General Studies
Equity	Total assets minus total liabilities	General Studies
Gross Debt	Total financial obligations of the company	General Studies
LP Loans	Long-Term Loans	General Studies
EBIT	Earnings Before Interest and Taxes	General Studies

EBITDA	Earnings before interest, taxes, depreciation, and amortization	General Studies
Net Profit	Profit after all costs, expenses, interest, and taxes	General Studies
ROE	Return on Equity	Soukal et al., 2023
P/VPA	Price to Book Value per Share	Sulistiani et al., 2022
FCO	Operating Cash Flow	Rodríguez-Masero & López-Manjón, 2020
FCI	Investing Cash Flow	General Studies
FCF	Free Cash Flow	General Studies
Accounts Receivable	Amount receivable for sales or services provided on credit	General Studies
Withdrawn Profits	Profit not distributed by the company and reinvested	General Studies
Market Value	Company Market Capitalization	General Studies
PL/AT	The ratio of Shareholders' Equity to Total Assets	General Studies
DT/PL	The ratio of Gross Debt to Equity	General Studies
DT/AT	The ratio of Gross Debt to Total Assets	General Studies
DLP/AT	The ratio of Long-Term Loans to Total Assets	Lombardo et al., 2022
DLP/PL	Ratio of Long-Term Loans to Equity	Lombardo et al., 2022
TO/PL (EBIT/PL)	Earnings Before Wealth Tax	Premalatha et al., 2023
RL/PL	Net Income to Equity Ratio	Premalatha et al., 2023
EBITDA/AT	EBITDA to Total Assets Ratio	Premalatha et al., 2023
EBIT/AT	EBIT to Total Assets Ratio	Premalatha et al., 2023
LL/AT	The ratio of Net Income to Total Assets	Premalatha et al., 2023
EBITDA/RL	EBITDA to Net Revenue Ratio	Premalatha et al., 2023
EBIT/RL	EBIT to Net Revenue Ratio	Premalatha et al., 2023
RL/AT	The ratio of Net Revenue to Total Assets	Premalatha et al., 2023
CGOL/AT	Net Operating Working Capital over Total Assets	General Studies
CGOL/RL	Net Operating Working Capital to Net Revenue	General Studies
CR/RL	Accounts Receivable to Net Revenue Ratio	General Studies
PC/RL	The ratio of Current Liabilities to Net Income	General Studies
Asset Growth	Percentage of change in Total Assets from one period to another	General Studies
Growth in Sales	Percentage change in Net Revenue from one period to another	General Studies
Change in ROE	Percentage change in Return on Equity from one period to another	General Studies
Change in P/VPA	Percentage change in Price over Book Value per Share from one period to another	General Studies
Z-Score	Altman Index: A combination of five different financial measures to estimate a company's risk of bankruptcy.	Barboza et al., 2017
Z-Score_Emergentes	Adapted for emerging markets, this score uses a modified formula better to reflect these environments' economic and financial conditions. It considers aspects such as adjusted liquidity, accumulated profitability, operating efficiency, and market capitalization regarding debt.	Barboza et al., 2017

Note: The description "general studies" was used when the variables were used in several different studies

The variables used encompass many characteristics of the companies. They can be distributed into six major groups: liquidity indicators, leverage indicators, profitability indicators, operational efficiency indicators, growth and change indicators, Altman's z-score indicator, and Altman's z-score for emerging countries. Barboza et al. (2017) eight techniques for predicting bankruptcy were tested, adopting more than 13,000 American companies with data from 1985 to 2013 as a sample. The authors argued that using Altman's Z-Score brought inaccurate predictions, even with more modern models, and opted for inserting other variables.

Following a similar idea, the Z-Score and Z-Score for emerging countries were inserted as predictor variables in the present study.

After defining the sample, models, and variables, information was collected on the dates of disclosure of the financial statements and the dates of change in the company's situation with the CVM. After obtaining these data, a binary response variable (target) was created, which received a value of 1 if the company had presented a bankruptcy situation in the eight quarters after the release of the results or a value of 0 if the company's situation was different. In this way, all models were trained to detect a possible bankruptcy up to eight quarters in advance.

Methodological procedures

The models were trained from 01/01/2000 to 03/31/2018. Records after these dates ended on 12/31/2023 and were used as a test. Thus, there is an approximate proportion of 76% for training and 24% for testing.

In addition to training using the standard system settings, the hyperparameters were optimized through Grid Search and Bayesian optimization. Overall, hyperparameter optimization did not improve metrics using Grid Search or Bayesian optimization. Thus, a few values were changed, as presented in Table 5.

Table 5

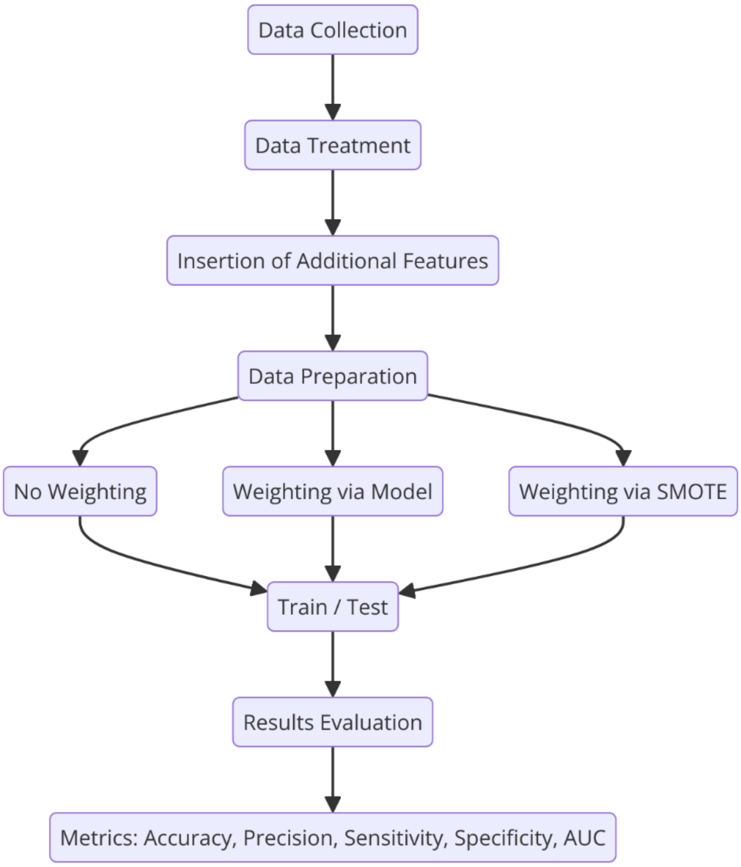
Hyperparameters used in the models

Model	Hyperparameters defined
XGBoost	No custom hyperparameters were used.
Random Forest	n_estimators=100
LightGBM	n_estimators=100
CatBoost	iterations=100
SVM	probability=True
Logistic Regression	max_iter=100 e solver='lbfgs

In predicting bankruptcy, there is a problem of class imbalance, which limits the performance of the models. In other words, the samples are composed chiefly of non-bankrupt companies. Much previous research has addressed the problem by applying resampling methods such as the SMOTE technique. However, resampling methods lead to such problems as increased noisy data and training time. An alternative to improve the bankruptcy prediction model is to use class weighting in the model parameters themselves, as this allows for dealing with the problem of class unbalance without the need for any resampling method (Yotsawat et al., 2023). In the present study, SMOTE and weighting via model parameters were used in both techniques. Figure 1 summarizes the methodology.

Figure 1

Research Framework



Source: Adapted from Muslim & Dasril, (2021)

Results

The descriptive analysis of the companies' financial data in Table 6, covering the period from 2000 to 2023 and 564 entities distributed in various industrial sectors, reveals a significant variability inherent to the economic and financial metrics. This variability can be primarily attributed to the length of the period analyzed and the sectoral heterogeneity of the companies included in the sample.

Table 6

Descriptive analysis of the study variables

Variable	Average	Desv. Pad.	Min.	25%	50%	75%	Max.
LI	1.27	12.06	-0.03	0.26	0.88	1.27	878.75
LC	2.98	23.03	0.00	1.28	2.40	2.98	1718.81
LS	2.53	22.92	-13.04	0.96	1.80	2.53	1718.81
LG	1.26	3.32	0.01	0.76	1.26	1.26	459.41
Current Assets	3091.25	7857.45	0.00	384.27	2687.86	3091.25	201926.00
Current liabilities	2139.15	5371.15	0.00	269.30	1922.17	2139.15	144169.00
Total Assets	10854.04	39435.47	0.00	986.75	8045.74	10854.04	1015142.00
Net Revenue	1531.28	4420.87	-10484.00	157.01	1167.48	1531.28	98260.00
Equity	4196.35	15910.49	-14148.75	337.13	2701.89	4196.35	362240.00
Gross Debt	3562.48	15498.64	0.00	187.07	2087.57	3562.48	506584.00
LP Loans	1321.66	2240.49	0.01	1321.66	1321.66	1321.66	70741.05

EBIT	190.72	1099.05	-44001.00	7.76	127.37	190.72	67836.00
EBITDA	344.81	1293.25	-31367.85	76.93	344.81	344.81	81162.00
Net Profit	82.03	910.46	-49724.00	1.84	68.18	82.03	60452.00
Profits withdrawn	34.83	801.57	-48523.00	31.55	34.83	34.83	54179.00
ROE	0.36	109.98	-7659.22	0.36	0.36	3.35	8694.15
P/VPA	1.62	8.49	-617.74	1.39	1.62	1.62	521.76
FCO	292.11	1156.42	-10243.31	39.06	292.11	292.11	46103.00
FCI	-220.72	896.42	-36973.41	-220.72	-220.72	-32.30	31097.00
CLF	-54.32	1163.38	-78945.00	-54.32	-54.32	-14.50	45331.33
Accounts Receivable	910.88	1874.54	0.00	116.32	904.02	910.88	43545.00
Market Value	6869.29	22408.84	0.01	448.75	5059.27	6869.29	897291.11
PL/AT	0.29	0.46	-9.87	0.29	0.29	0.46	1.00
DT/PL	1.14	8.41	-514.75	0.38	1.14	1.14	421.50
DT/AT	0.29	0.19	0.00	0.22	0.29	0.32	3.35
DLP/AT	0.12	0.08	0.00	0.12	0.12	0.12	3.34
DLP/PL	0.39	1.40	-54.67	0.39	0.39	0.39	106.25
TO/PL	0.02	2.05	-279.63	0.02	0.02	0.05	57.39
RL/PL	0.66	4.35	-139.70	0.27	0.66	0.66	302.68
EBITDA/AT	0.02	0.27	-26.59	0.02	0.02	0.02	30.78
EBIT/AT	0.01	0.35	-40.39	0.01	0.01	0.02	30.69
LL/AT	0.00	0.33	-8.97	0.00	0.00	0.01	46.50
EBITDA/RL	4.69	403.59	-7503.36	0.17	4.69	4.69	61087.06
EBIT/RL	2.80	403.56	-7529.55	0.06	0.20	2.80	61027.22
RL/AT	0.19	0.11	-1.51	0.13	0.19	0.19	2.04
CGOL/AT	0.09	0.34	-9.87	0.07	0.09	0.19	1.00
GOL/RL	-1.42	620.94	-72931.00	-1.42	-0.19	1.00	33965.32
CR/RL	2.58	32.71	-262.39	0.67	1.32	2.58	2565.83
PC/RL	14.55	576.92	-36978.72	1.32	2.91	14.55	74345.00
Asset Growth	318073.66	33212468.04	-99.92	0.00	4.17	318073.66	5077784410.87
Growth in Sales	1159.93	114276.45	-14274.38	0.00	10.35	1159.93	17453500.00
Change in ROE	3.83	10982.81	-933616.63	-45.94	3.83	3.83	884448.27
Change in P/VPA	22.98	1238.99	-24246.55	0.00	22.98	22.98	137975.17
Z-Score	5.62	48.88	-808.96	5.40	5.62	5.62	4289.87
Z-Score_Emergentes	6.96	11.32	-271.10	6.96	6.96	6.96	664.49
close_adj_1d	900.72	10534.16	0.00	7.73	468.43	900.72	474851.92
close_adj_30d	904.21	10422.60	0.00	7.90	904.21	904.21	483919.44
close_adj_60d	896.17	10592.68	0.00	8.01	798.08	896.17	535222.54
close_adj_90d	897.74	10606.11	0.00	8.00	644.72	897.74	502531.73

Specifically, the wide dispersion evidenced by the high standard deviations about the averages observed in variables such as "Current Assets," "Current Liabilities," and "Market Value" suggests a substantial diversity in the financial conditions of companies. This dispersion is probably influenced by different economic cycles, which include periods of expansion and recession, directly reflecting on the financial performance of companies over time.

Notably, the extreme values in 'Net Income' and 'EBITDA' serve as crucial indicators of extraordinary events or high economic volatility. These events, which could be linked to specific economic crises, abrupt regulatory changes, or disruptive innovations, have disproportionately affected specific industries, significantly impacting the companies' finances (Santos & Peixoto, 2023; Haslam, 2017).

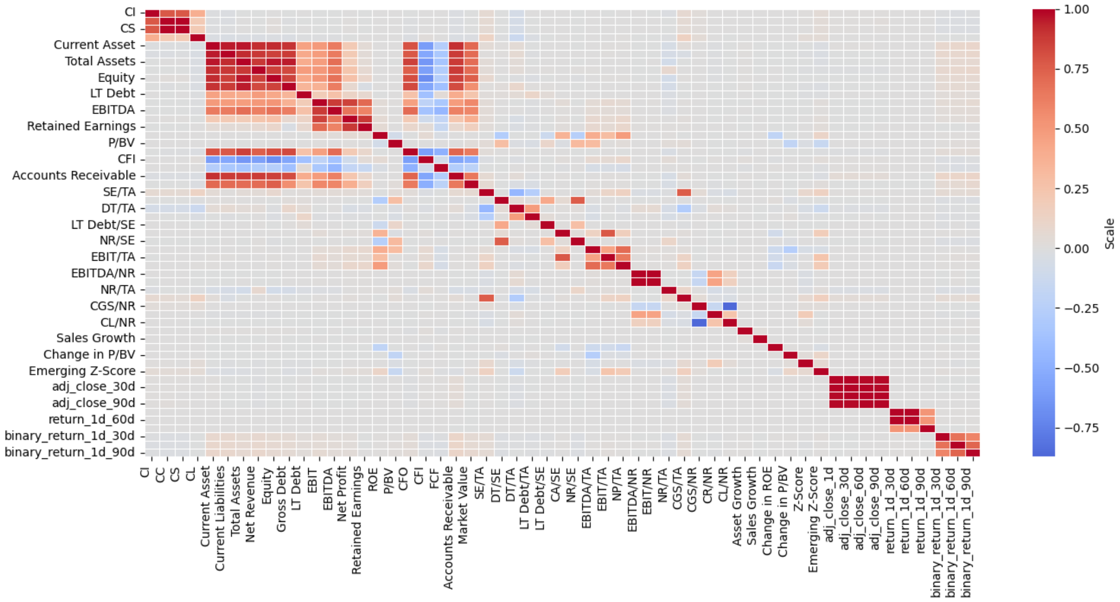
The data's distribution by quartiles shows that a large proportion of values for many financial metrics are concentrated below the average. This pattern may suggest that a considerable number of companies faced sustained financial challenges or operated at reduced levels of economic efficiency during the period considered.

Finally, the diversity of sectors, as presented in the methodological section through Table 2, with significant representation in areas such as "Cyclical Consumption," "Industrial Goods," and "Public Utility," and more miniature representations in sectors such as "Hotels and Restaurants" and "Chemicals," suggests that sectoral specificities play a crucial role in the observed financial dynamics. Sector-specific economic conditions, including sensitivity to macroeconomic and regulatory factors, are critical determinants of recorded financial fluctuations (Khmeleva et al., 2023)

This comprehensive analysis underscores the importance of considering the extended time context and sectoral heterogeneity. Such considerations are vital when assessing firms' financial health and developing investment strategies based on robust economic and financial assessments.

Figure 1

Correlations heat map



The heat map illustrates the correlation matrix between several financial variables, focusing only on their linear relationships. A strong positive correlation between variables that reflect the company's economic structure, such as "Total Assets" and "Shareholders' Equity," shows that significant assets are often associated with robust shareholders' equity. However, variables related to period-adjusted closing prices, such as "close_adj_30d" and "close_adj_60d", among others, show little or no linear correlation with traditional financial metrics. This suggests that movements in closing prices are not directly linked to the company's static financial conditions. Such an observation is important as it indicates that external or market factors can significantly impact stock prices that are not directly noticeable through conventional economic measures. This linear correlation analysis is a starting point for further investigations, possibly exploring models that capture more complex nonlinear or dynamic relationships.

Table 7

Models Analysis Metrics for Predicting Bankruptcy 8 Quarters Ahead

No weighting	XGB	RF	LGB	CAT	SVM	RL
Accuracy	0.95	0.93	0.98	0.98	0.42	0.88
Precision	0.09	0.06	0.23	0.21	0.01	0.03
Sensitivity (Recall)	0.84	0.89	0.84	0.74	0.89	0.68
Specificity	0.95	0.93	0.98	0.98	0.42	0.88
AUC	0.94	0.94	0.94	0.91	0.81	0.85
Best Cutoff Point	0.00	0.06	0.00	0.02	0.02	0.03

SMOTE Weighting	XGB	RF	LGB	CAT	SVM	RL
Accuracy	0.91	0.96	0.85	0.99	0.77	0.74
Precision	0.05	0.12	0.03	0.53	0.02	0.02
Sensitivity (Recall)	0.95	0.84	0.95	0.84	0.79	0.84
Specificity	0.91	0.96	0.85	1.00	0.77	0.74
AUC	0.97	0.94	0.96	0.93	0.84	0.86
Best Cutoff Point	0.00	0.11	0.00	0.52	0.18	0.48

Note: XGB: XGBoost, RF: Random Forest; LGB: Light Gradient Boosting Machine; CAT: CatBoost; SVM: Support Vector Machine; RL: Logistic Regression

Analysis of performance metrics for different machine learning models applied to business failure prediction eight quarters ahead reveals the effectiveness of each technique in specific contexts. This study considers a diverse range of algorithms, both in non-weighted and SMOTE-weighted scenarios, as detailed in Table 7. The class weighting within the model parameters was not included in the table as it yielded results similar to those achieved through SMOTE weighting.

In the unweighted scenarios, it is observed that the RF and LGB models present superior performance in terms of accuracy and specificity, both reaching 0.98 in these metrics, which indicates a high capacity to classify both bankruptcies and non-bankruptcies correctly. On the other hand, the accuracy of these models is relatively low (0.23 for RF and 0.21 for LGB), suggesting a considerable proportion of false positives among the failure predictions. Notably, all models except MVS show a high AUC (Area Under the ROC Curve) value, with 0.94, indicating an excellent overall balance between sensitivity and specificity.

When applied to SMOTE weighting, which aims to balance unbalanced classes through the oversampling technique, there is a significant improvement in several metrics for some models. The CAT model stands out with an accuracy of 0.99 and a perfect specificity of 1.00, in addition to a precision of 0.53, significantly higher than the other models. In addition, the AUC for CAT shows an excellent performance, with 0.93. This result suggests that SMOTE weighting may be particularly beneficial for gradient boosting-based algorithms, as evidenced by increased accuracy and AUC for LGB and CAT.

In contrast, despite benefiting from class adjustment with SMOTE weighting, the SVM model continues to underperform compared to boosting methods, standing out as less suitable for this type of failure prediction when the data is highly unbalanced.

This detailed analysis of models in two distinct data weighting contexts not only underscores the importance of choosing the appropriate model for specific financial forecasts but also demonstrates the effectiveness of the SMOTE technique in improving the performance of models on unbalanced data. These results are relevant to the bankruptcy prediction literature and predictive practices in corporate finance, offering a robust basis for future investigations on applying machine learning techniques in long-term financial forecasting.

Given the results presented when CatBoost was used, we chose to perform a more in-depth analysis, which is given in Table 8, analyzing the result at different cutoff points.

Table 8

Estimates using Catboost with probability cutoffs ranging from 30% to 80%

Metrics	CAT30	CAT40	CAT50	CAT60	CAT70	CAT80
Accuracy	0.99529	0.99588	0.99647	0.99706	0.99735	0.99765
Precision	0.56000	0.61905	0.70588	0.80000	0.91667	1.00000
Sensitivity (Recall)	0.73684	0.68421	0.63158	0.63158	0.57895	0.57895
Specificity	0.99675	0.99763	0.99852	0.99911	0.99970	1.00000
AUC	0.95665	0.95665	0.95665	0.95665	0.95665	0.95665
F1 Score	0.63636	0.65000	0.66667	0.70588	0.70968	0.73333

Table 8 shows the incremental performance of the CatBoost model in predicting bankruptcies with the range of cutoff points from 30% to 80%. This analysis is essential to understand how changing decision thresholds affect the model's accuracy, precision, sensitivity (recall), specificity, and AUC. As the cutoff points increase, a consistent increase in accuracy and specificity is observed, indicating an improvement in the model's ability to identify non-bankrupt companies correctly. This capacity for discernment intensifies, reaching a specificity of 100% at the highest cutoff point, evidencing a perfect classification of non-failures.

At the same time, the model's accuracy also progressively improves with the increase in the cutoff point, starting at 56% and culminating at 100% at the cutoff end of 80%. This gain in accuracy suggests that the model becomes highly reliable in its predictions of bankruptcies, aligning the security in the prediction with the increase in the threshold. However, this adjustment comes with a reduction in sensitivity, which declines from 73.684% to 57.895% as the cutoff point is raised, reflecting the model's decreasing ability to capture all real failures, a consequence of the compromise between reducing false positives and increasing false negatives.

This behavior of the CatBoost model contrasts and complements the observations of other studies that employed different methods of prediction and analysis of failures, such as those described by Shah et al. (2022) and Sulistiani et al. (2022). While the first authors highlighted the effectiveness of the Random Forest model in classification based on diverse attributes, Sulistiani et al. (2022) found benefits in using machine learning techniques to mitigate overfitting and imbalance in the data. As presented in this study, using varied decision thresholds in CatBoost reflects a strategic approach to maximize accuracy while managing the trade-off between sensitivity and specificity.

The constant AUC of 95.665% was already expected, given that it does not depend on the cutoff point. Still, its high value reinforces CatBoost's overall ability to maintain a high level of discrimination between bankruptcies and non-bankruptcies. This indicator aligns with the high AUC values observed in studies such as those by Lombardo et al. (2022), in which different machine-learning models were explored.

Additionally, the improvement in F1 Score values with the increase in cutoff points highlights the refinement in the model configuration. This emphasizes that careful threshold calibration can improve the model's overall performance, allowing for an optimal balance between avoiding erroneous predictions and not omitting crucial failure identifications.

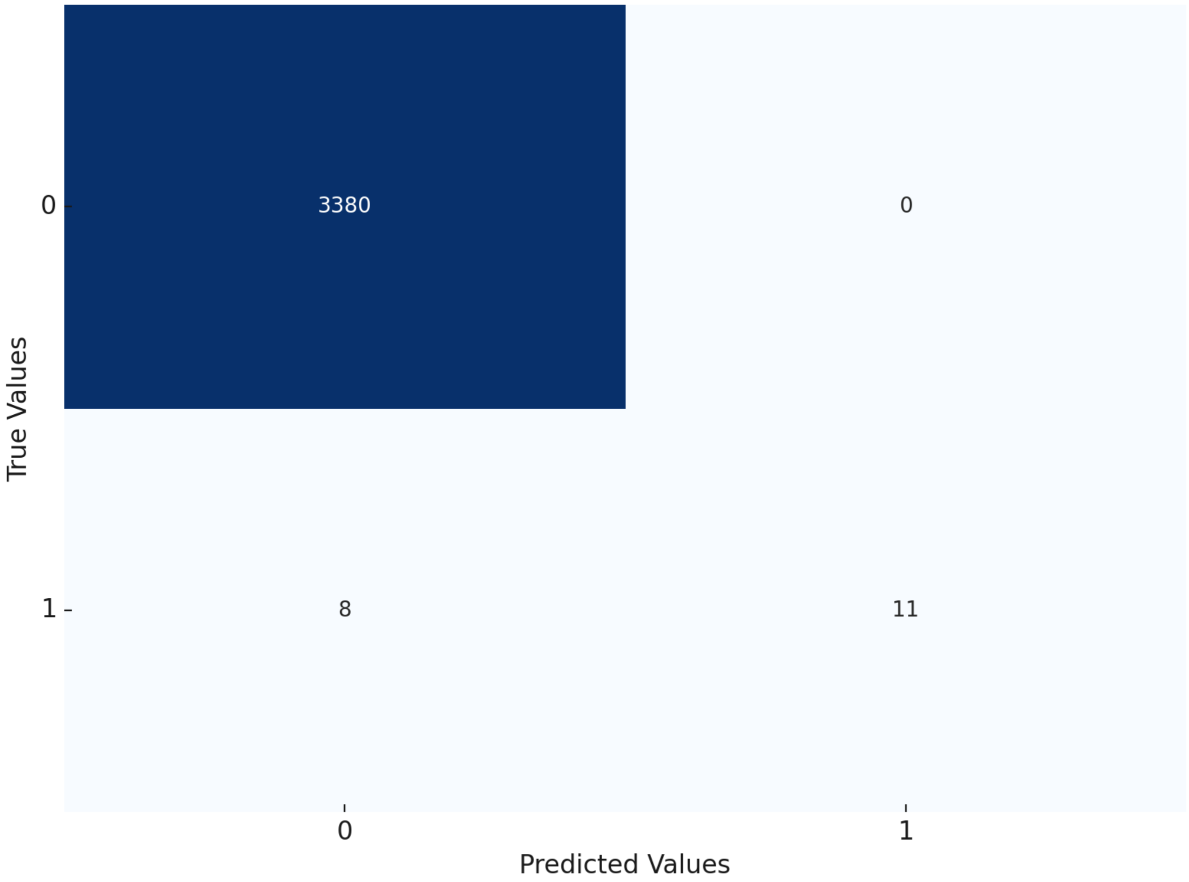
Therefore, this study's results not only corroborate previous research findings, indicating the effectiveness of machine learning techniques in predicting bankruptcies but also expand the understanding of how threshold manipulation can be decisive in optimizing the performance of predictive models in complex financial contexts.

As pointed out by Kristanti and Dhaniswara (2023) and by Sun, Liu, and Sima (2020), in addition to adjusting the cutoff point and applying oversampling, we tried to improve the metrics using more appropriate hyperparameters than the predefined ones. For this, grid search and Bayesian optimization were used. Still, besides consuming many machine resources, this process did not produce results superior to those of the standard system configuration.

The analysis reveals that higher cutoff points can increase accuracy and specificity and reduce sensitivity acceptance. Therefore, choosing the optimal cutoff point must be guided by carefully evaluating the costs associated with each classification error type, which must align with the strategic priorities of the model's use.

Figure 2

CAT80 Confusion Matrix



In Figure 2, the confounding matrix generated by the CatBoost model with a cutoff point of 80% (CAT80) reveals a distinct performance in predicting bankruptcies, standing out for a high rate of true negatives (TN = 3,380) and a total absence of false positives (FP). This characteristic of the model is comparable, for example, to the study by Sulistiani et al. (2022), in which the application of balancing techniques such as BSM aimed to minimize the incidence of false positives in bankruptcy prediction models. However, CatBoost's approach, focusing on high specificity and a perfect non-bankruptcy identification rate, demonstrates the effectiveness of high decision thresholds to avoid the risk of critical misclassifications in financial contexts.

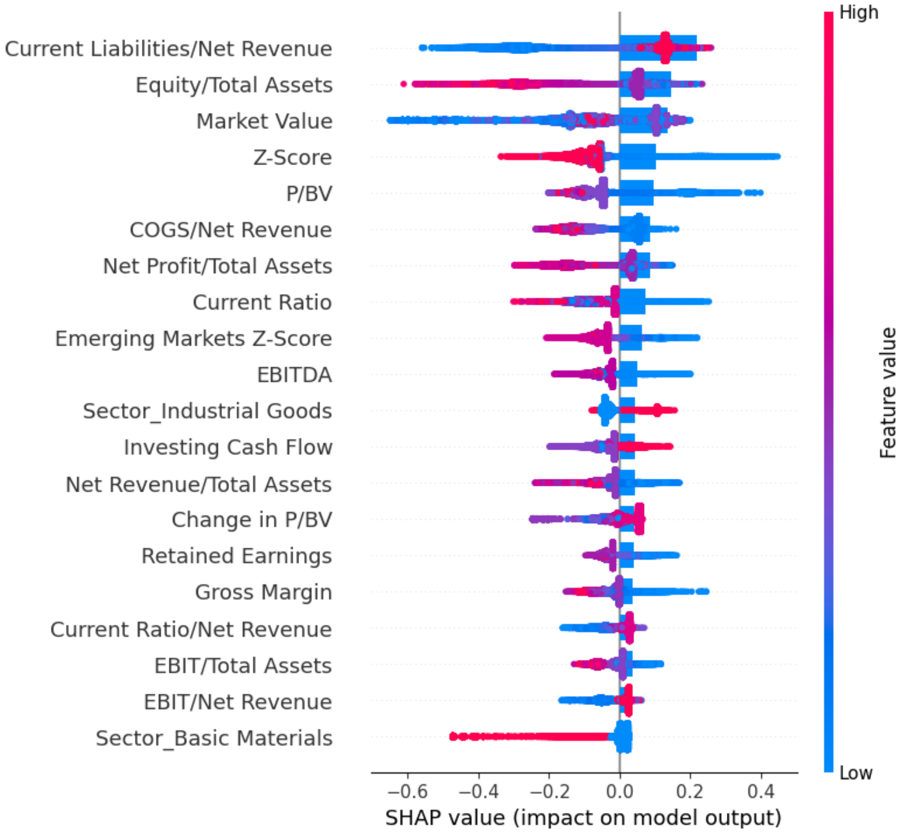
On the other hand, CAT80's sensitivity is moderate since it identifies fewer actual failures (TP = 11) with eight false negatives (FN). This detection rate is similar to the challenges faced by Shah et al. (2022), who also observed limitations in their models regarding capturing all real failures. However, these studies emphasize the importance of a balance between sensitivity and specificity, which CAT80 sacrifices in favor of the latter.

Additionally, the high AUC (95.665%) reflects a capacity for discrimination. This finding resonates with the results obtained by Lombardo et al. (2022), in which different machine learning models demonstrated strong performance in distinguishing between failures and non-failures through similar AUC metrics. The consistency of CAT80 in keeping this metric high suggests that, as noted in the study by Lombardo et al. (2022), machine learning techniques can offer information relevant to financial practice, even under conditions of strict parameterization.

The expanded discussion of these comparisons illuminates the complexity and nuances of applying predictive models in finance, especially in critical tasks such as bankruptcy prediction. While the CAT80 model demonstrates remarkable effectiveness in ensuring that forecasts are reliable, the need for more comprehensive failure detection, as shown in the studies above, underscores the importance of carefully choosing thresholds and methods to optimize the balance between different performance metrics.

Figure 3

SHAP values of variables



The detailed analysis of the SHAP values, illustrated in Figure 3, reveals the implications of the top 20 traits, out of 49 traits, on the model's ability to predict bankruptcies. Financial and sectoral variables stand out, whose varied influences are essential to understanding the predictive behavior of the system in a business context. Notably, 'Current Liabilities/Net Revenue' emerges as the variable with the most significant impact, presenting a positive correlation with the probability of bankruptcy at high values. This result echoes the findings of Sulistiani et al. (2022), which associated high liabilities with indicators of financial stress. High current liabilities relative to net revenue signal financial strain, thus increasing the likelihood of insolvency.

'Equity/Total Assets' is another critical variable, negatively correlated with the probability of bankruptcy at high values. A higher equity to total assets ratio suggests a more robust financial foundation, reducing the risk of bankruptcy. This supports the perspective that companies with substantial equity bases are perceived as more financially stable. Similarly, 'Market Value' shows a significant impact, with higher market valuations indicating a lower risk of financial difficulties. This finding is consistent with Wang et al. (2023), who emphasized the importance of market valuations in the financial sustainability of companies.

Sector dummy variables illustrate how classification in specific sectors affects bankruptcy predictions. For instance, the 'Sector_Basic Materials' and 'Sector_Industrial Goods' exhibit varied impacts on bankruptcy risk, indicating that companies in these sectors have unique financial characteristics. This observation aligns with the analysis by Thilakarathna et al. (2022), which highlighted the influence of sector-specific factors on financial stability.

Operational performance indicators such as 'EBIT/Net Revenue' and 'EBIT/Total Assets' further confirm the significance of operational efficiency in predicting bankruptcies. High values of these indicators are associated with a reduced risk of bankruptcy, reinforcing the findings of Premalatha et al. (2023). Efficient operations enhance a company's financial health, lowering the likelihood of financial distress.

This SHAP value chart highlights the relative importance of each trait and sheds light on the complex interactions between financial and industry indicators in predicting bankruptcies. It emphasizes how these findings can inform corporate policies and strategic decisions, improving various industries' risk management and resource allocation. This analysis provides a basis for future research on predictive modeling in finance, promoting a more sophisticated approach to elaborating predictive models of bankruptcies.

Limitations and future work

Although this study provides relevant information on predicting bankruptcies using the CatBoost model, some limitations must be considered. First, historical financial data may not fully capture the effects of unexpected economic events or global financial crises. In addition, using predominantly quantitative data can omit critical qualitative factors that affect companies' financial health, such as changes in management or the regulatory environment. Future research could address these limitations by incorporating scenario analysis and qualitative data to enrich predictive models.

One trend observed in the literature on bankruptcy forecasting is the increasing diversification of the data sources employed, moving beyond traditional financial indicators to include heterogeneous data from multiple sources, such as news texts, annual reports, and macroeconomic indicators (Soukal et al., 2023). This evolution reflects a paradigmatic shift in the field of bankruptcy prediction, as demonstrated by Qu et al. (2019), who discussed the increasing use of advanced techniques such as CNN to analyze such data.

This move toward richer data and more complex models suggests a promising avenue for future research, particularly in how this new data can be integrated effectively to improve the accuracy of predictions. Wang et al. (2023) have already demonstrated the effectiveness of engineering advanced features in financial records, a method that could be expanded to incorporate these new data types.

In addition, the transition from simple statistical techniques to hybrid models and set methods, as explored by El Maldou et al. (2023), highlights the ongoing search for models that address data complexity and improve the robustness of forecasts by integrating multiple predictive approaches. Applying these advanced techniques to heterogeneous data represents a fertile field for future research, with the potential to develop more profound studies and more reliable predictions of bankruptcies.

Thus, future work should continue to explore the expansion of the types of data used in predicting bankruptcies and continuous innovation in modeling techniques to address existing limitations and fully exploit the potential of complex data in predicting bankruptcies. These advances could help mitigate financial risks and strengthen global economic stability by aligning with emerging financial research trends and diverse stakeholders' practical requirements.

Conclusion

This study explored applying six machine learning models to predict bankruptcies in Brazilian companies listed on B3, emphasizing the importance of adjusting and adapting these models to specific economic contexts. A comprehensive analysis identified several key findings, including Random Forest, XGBoost, LightGBM, CatBoost, SVM, and Logistic Regression.

Firstly, machine learning models, particularly those utilizing boosting techniques like CatBoost, demonstrated significant effectiveness in bankruptcy prediction. CatBoost, combined with the SMOTE technique for class balancing, achieved superior performance metrics, including an accuracy of 99% and a specificity of 100% at specific cutoff points. This underscores the potential of boosting algorithms in handling imbalanced datasets and improving prediction accuracy.

Secondly, the SHAP value analysis revealed the critical importance of specific financial and sectoral variables in predicting bankruptcies. Variables such as market value, current liabilities, and gross debt were significant predictors, aligning with previous studies highlighting financial stress indicators. Additionally, sectoral dummy variables indicated that industry-specific factors significantly impact bankruptcy prediction, corroborating findings that sector characteristics play a crucial role in financial stability.

Thirdly, while hyperparameter optimization through Grid Search and Bayesian optimization was explored, it did not consistently improve model metrics. This suggests that default model parameters can be robust enough for practical applications, though customization remains essential for specific datasets and contexts.

Fourthly, addressing class imbalance was crucial for enhancing model performance. Applying the SMOTE technique proved effective in improving the detection of bankruptcies. This approach mitigates the limitations of highly imbalanced datasets, a common challenge in predicting financial distress.

Finally, operational performance indicators such as EBIT/RL and RL/AT were validated as essential components in bankruptcy prediction, emphasizing the role of operational efficiency in maintaining financial health. This finding aligns with existing literature that underscores the importance of efficient operations in preventing financial distress.

In conclusion, this study contributes valuable insights into applying machine learning models for bankruptcy prediction in the Brazilian market. The findings highlight the effectiveness of advanced machine-learning techniques and the significance of financial and sector-specific variables. Future research should consider incorporating qualitative data and further exploring these techniques in different economic and industrial contexts to enhance the robustness and applicability of bankruptcy prediction models.

References

Baldissera, J. F., Fiirst, C., Rovaris, N. R., & Dall'Asta, D. (2020). Estrutura de Capital em Empresas Brasileiras Listadas na B3 nos Anos Antecedentes ao Pedido de Falência ou Recuperação Judicial. *Revista Contabilidade e Controladoria*, *11*(2). https://doi.org/10.5380/rcc.v11i2.67196

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

Ben-Hur, A., & Weston, J. (2010). *A User's Guide to Support Vector Machines* (pp. 223–239). https://doi.org/10.1007/978-1-60327-241-4_13

Bittencourt, W. R., & Albuquerque, P. H. M. (2020). Evaluating company bankruptcies using causal forests. *Revista Contabilidade & Finanças*, *31*(84), 542–559. https://doi.org/10.1590/1808-057x202010360

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *42*(8), 785–794. https://doi.org/10.1145/2939672.2939785

Chen, T., He, T., & Benesty, M. (2018). xgboost : eXtreme Gradient Boosting. *R Package Version 0.71-2*, 1–4.

Chen, Z., Chen, W., & Shi, Y. (2020). Ensemble learning with label proportions for bankruptcy prediction. *Expert Systems with Applications*, *146*. https://doi.org/10.1016/j.eswa.2019.113155

Dos Santos, R. R., & Peixoto, F. M. (2023). Financial distress e governança corporativa: um estudo no mercado de capitais brasileiro. *Revista de Gestão e Secretariado (Management and Administrative Professional Review)*, *14*(11), 20158–20201. https://doi.org/10.7769/gesec.v14i11.3172

El Madou, K., Marso, S., El Kharrim, M., & El Merouani, M. (2023). Evolutions in machine learning technology for financial distress prediction: A comprehensive review and comparative analysis. *Expert Systems*. https://doi.org/10.1111/exsy.13485

Ferren, F., & Kurniadi, F. I. (2023). We are enhancing Bankruptcy Prediction with Feature Selection in the AdaBoost Algorithm. *2023 10th International Conference on ICT for Smart Society (ICISS)*, 1–4. https://doi.org/10.1109/ICISS59129.2023.10291988

Giriūniene, G., Giriūnas, L., Morkunas, M., & Brucaite, L. (2019). A Comparison on Leading Methodologies for Bankruptcy Prediction: The Case of the Construction Sector in Lithuania. *Economies*, *7*(3), 82. https://doi.org/10.3390/economies7030082

Guimarães, A. L. de S., & Alves, W. O. (2009). Prevendo a insolvência de operadoras de planos de saúde. *Revista de Administração de Empresas*, *49*(4), 459–471. https://doi.org/10.1590/S0034-75902009000400009

Guimarães, P. R. F., & Resende Filho, M. de A. (2018). Uma aplicação do modelo de regressão logística na previsão de falência empresarial no Brasil. *R. Bras. Eco. de Emp*, *18*(2), 21–42.

Haslam, C. (2017). International Financial Reporting Standards (IFRS): Stress testing in financialized reporting entities. *Accounting, Economics and Law*, *7*(2), 105–108. https://doi.org/10.1515/ael-2017-0016

Heo, J., & Yang, J. Y. (2014). AdaBoost based bankruptcy forecasting of Korean construction companies. *Applied Soft Computing Journal*, *24*, 494–499. https://doi.org/10.1016/j.asoc.2014.08.009

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *2017-Decem*(Nips), pp. 3147–3155.

Khalil, R. A. B., & Bakar, A. A. B. U. (2023). A Comparative Study of Deep Learning Algorithms in Univariate and Multivariate Forecasting of the Malaysian Stock Market. *Sains Malaysiana*, *52*(3), 993–1009. https://doi.org/10.17576/jsm-2023-5203-22

Khmeleva, G. A., Semenychev, V. K., Korobetskaya, A. A., Kurnikova, M. V., Fedorenko, R., & Tóth, B. I. (2023). Comparative Research of Internal and Border Regions: Analyzing the Differences in the Cyclical Dynamics of Industries for Industrial Policy and Territorial Development. *Economies*, *11*(3). https://doi.org/10.3390/economies11030089

Kovacova, M., Kliestik, T., Valaskova, K., Durana, P., & Juhaszova, Z. (2019). Systematic review of variables applied in bankruptcy prediction models of Visegrad group countries. *Oeconomia Copernicana*, *10*(4), 743–772. https://doi.org/10.24136/oc.2019.034

Kristanti, F. T., & Dhaniswara, V. (2023). The Accuracy of Artificial Neural Networks and Logit Models in Predicting the Companies' Financial Distress. *Journal of Technology Management & Innovation*, *18*(3), 42–50. https://www.jotmi.org/index.php/GT/article/view/4149

Letkovský, S., Jenčová, S., & Vašaničová, P. (2024). Is Artificial Intelligence Really More Accurate in Predicting Bankruptcy? *International Journal of Financial Studies*, *12*(1). https://doi.org/10.3390/ijfs12010008

Liashenko, O., Kravets, T., & Kostovetskyi, Y. (2023). Machine Learning and Data Balancing Methods for Bankruptcy Prediction. *Ekonomika* , *102*(2), 28–46. https://doi.org/10.15388/Ekon.2023.102.2.2

Lombardo, G., Pellegrino, M., Adosoglou, G., Cagnoni, S., Pardalos, P. M., & Poggi, A. (2022). Machine Learning for Bankruptcy Prediction in the American Stock Market: Dataset and Benchmarks. *Future Internet*, *14*(8), 244. https://doi.org/10.3390/fi14080244

Lott, V. F., Tenenwurcel, D. R., & Camargos, M. A. de. (2021). Determinantes do endividamento de empresas brasileiras listadas na B3 com e sem risco de insolvência. *Revista de Administração Da UFSM*, *14*(1), 79–99. https://doi.org/10.5902/1983465933892

Máté, D., Raza, H., & Ahmad, I. (2023). Comparative Analysis of Machine Learning Models for Bankruptcy Prediction in the Context of Pakistani Companies. *Risks*, *11*(10), 176. https://doi.org/10.3390/risks11100176

Muslim, M. A., & Dasril, Y. (2021). Company bankruptcy prediction framework based on the most influential features using XGBoost and stacking ensemble learning. *International Journal of Electrical and Computer Engineering*, *11*(6), 5549–5557. https://doi.org/10.11591/ijece.v11i6.pp5549-5557

Pereira, V. S., & Martins, V. F. (2016). Estudos de previsão de falências – uma revisão das publicações internacionais e brasileiras de 1930 a 2015. *Revista Contemporânea de Contabilidade*, *12*(26), 163. https://doi.org/10.5007/2175-8069.2015v12n26p163

Platt, H. D., & Platt, M. B. (2002). Predicting corporate financial distress: Reflections on choice-based sample bias. *Journal of Economics and Finance*, *26*(2), 184–199. https://doi.org/10.1007/BF02755985

Prado, J. W. do, Carvalho, F. de M., Benedicto, G. C. de, & Lima, A. L. R. (2019). Analysis of credit risk faced by public companies in Brazil: an approach based on discriminant analysis, logistic regression and artificial neural networks. *Estudios Gerenciales*, *35*(153), 347–360. https://doi.org/10.18046/j.estger.2019.153.3151

Premalatha, G., Priyanka, R., & Chaitya, K. (2023). Feature selection for predicting bankruptcy: Comparative analysis. *2023 Fifth International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, 1–5. https://doi.org/10.1109/ICECCT56650.2023.10179633

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2017). *CatBoost: unbiased boosting with categorical features*. *Section 4*, 1–23. http://arxiv.org/abs/1706.09516

Qu, Y., Quan, P., Lei, M., & Shi, Y. (2019). Review of bankruptcy prediction using machine learning and deep learning techniques. *Procedia Computer Science*, *162*, 895–899. https://doi.org/10.1016/j.procs.2019.12.065

Rodríguez-Masero, N., & López-Manjón, J. (2020). The Usefulness of Operating Cash Flow for Predicting Business Bankruptcy in Medium-Sized Firms. *Review of Business Management*, *22*(4), 917–931. https://doi.org/10.7819/rbgn.v22i4.4079

Securities and Exchange Commission. (2023). General Register of Publicly-Held Companies and Registered Companies. Retrieved from https://dados.cvm.gov.br/dados/CIA_ABERTA/CAD/DADOS/cad_cia_aberta.csv

Shah, J., Rao, B., Mehta, Y., & Kurhade, S. (2022). Predicting Bankruptcy and Suggesting Improvements on Financial Attributes using Machine Learning Models. *2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 807–812. https://doi.org/10.1109/ICESC54411.2022.9885647

Shetty, S., Musa, M., & Brédart, X. (2022). Bankruptcy Prediction Using Machine Learning Techniques. *Journal of Risk and Financial Management*, *15*(1). https://doi.org/10.3390/jrfm15010035

Soukal, I., Mačí, J., Trnková, G., Svobodova, L., Hedvičáková, M., Hamplova, E., Maresova, P., & Lefley, F. (2023). A state-of-the-art appraisal of bankruptcy prediction models focussing on the field's core authors: 2010–2022. *Central European Management Journal*. https://doi.org/10.1108/CEMJ-08-2022-0095

Sulistiani, I., Widodo, & Nugraheni, M. (2022). Comparison of Bankruptcy Prediction Models Using Support Vector Machine and Artificial Neural Network. *2022 11th Electrical Power, Electronics, Communications, Controls and Informatics Seminar (EECCIS)*, 316–321. https://doi.org/10.1109/EECCIS54468.2022.9902935

Sun, X., Liu, M., & Sima, Z. (2020). A novel cryptocurrency price trend forecasting model based on LightGBM. *Finance Research Letters*, *32*(November 2018), 101084. https://doi.org/10.1016/j.frl.2018.12.032

Syafei, R. M., & Efrilianda, D. A. (2023). Machine Learning Model Using Extreme Gradient Boosting (XGBoost) Feature Importance and Light Gradient Boosting Machine (LightGBM) to Improve Accurate Prediction of Bankruptcy. *Recursive Journal of Informatics*, *1*(2), 64–72. https://doi.org/10.15294/rji.v1i2.71229

Thanh-Long, N., Minh, T.-, & Hong-Chuong, L. (2022). A Back Propagation Neural Network Model with the Synthetic Minority Over-Sampling Technique for Construction Company Bankruptcy Prediction. *International Journal of Sustainable Construction Engineering and Technology*, *13*(3), 68–79. https://doi.org/10.30880/ijscet.2022.13.03.007

Thilakarathna, C., Dawson, C., & Edirisinghe, E. (2022). Using Financial Ratios with Artificial Neural Networks for Bankruptcy Prediction. *2022 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, pp. 55–58. https://doi.org/10.1109/ICAICA54878.2022.9844640

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory*. Springer, New York. https://doi.org/10.1007/978-1-4757-3264-1

Wang, X., Kräussl, Z., Zurad, M., & Brorsson, M. (2023). Effective Automatic Feature Engineering on Financial Statements for Bankruptcy Prediction. *2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, 1–8. https://doi.org/10.1109/ICECCME57830.2023.10252608

Yotsawat, W., Phodong, K., Promrat, T., & Wattuya, P. (2023). Bankruptcy prediction model using cost-sensitive extreme gradient boosting in the context of imbalanced datasets. *International Journal of Electrical and Computer Engineering (IJECE)*, *13*(4), 4683. https://doi.org/10.11591/ijece.v13i4.pp4683-4691

Yu, Q., Miche, Y., Séverin, E., & Lendasse, A. (2014). Bankruptcy prediction using Extreme Learning Machine and financial expertise. *Neurocomputing*, *128*, 296–302. https://doi.org/10.1016/j.neucom.2013.01.063