Forecasting Corporate Bankruptcy: A Comparative Study of Statistical and Machine Learning Models

Ritika Vaidya, Anokhi Desai, Sri Venkata Sai Lohit Ganduri
New York University
rv2460@nyu.edu, adp9797@nyu.edu, sg8874@nyu.edu

May 7, 2025

Abstract

This paper compares statistical and machine learning models in predicting corporate bankruptcy using financial data from 181 U.S. manufacturing firms. We evaluate logistic regression, Lasso, Ridge, Random Forest, and Extreme Gradient Boosting (XGBoost) based on accuracy, sensitivity, specificity, and AUC. While logistic regression offers interpretability, it is outperformed by regularized and ensemble methods. XGBoost achieves the highest out-of-sample AUC (0.8693), highlighting its effectiveness in capturing complex patterns. The results suggest that while linear models aid interpretation, ensemble learning provides superior predictive performance for early detection of financial distress.

1 Introduction

Bankruptcy signifies a firm's inability to meet its debt obligations, often resulting in severe financial and operational consequences. Accurately predicting corporate bankruptcy has long been a critical concern for investors, creditors, and regulators. Early detection of financially distressed firms allows for timely intervention, more efficient capital allocation, and reduced systemic risk. Classical statistical models—such as Altman's Z-Score (Altman, 1968), logistic regression (Ohlson, 1980), and multivariate discriminant analysis (Altman, Haldeman & Narayanan, 1977)—have historically guided risk evaluation. However, these models are limited by assumptions of linearity, normality, and homoscedasticity, and often struggle with multicollinearity and high-dimensional input data.

Machine learning (ML) techniques have gained prominence in financial prediction tasks due to their ability to model nonlinear relationships, handle large and noisy datasets, and adapt to complex data structures (Lessmann et al., 2015; Huang et al., 2004). Regularized regression methods such as Lasso (Tibshirani, 1996) and Ridge (Hoerl & Kennard, 1970) provide a principled approach to mitigate overfitting and multicollinearity through penalization, making them especially suitable for financial ratio analysis. Lasso also facilitates variable selection by shrinking some coefficients to zero, enhancing interpretability—an advantage in financial settings where understanding key predictors is vital.

This study implements and compares a suite of models—logistic regression, Lasso, Ridge, Random Forest (Breiman, 2001), and Extreme Gradient Boosting (Chen & Guestrin, 2016)—to predict corporate bankruptcy using financial data from U.S. manufacturing firms. We evaluate each model using performance metrics such as accuracy, sensitivity, specificity, and area under the ROC curve (AUC). To contextualize our approach, we begin by reviewing key models in the evolution of bankruptcy prediction—from classical statistical techniques to modern machine learning approaches.

2 Literature Review

Altman, Haldeman, and Narayanan (1977) advanced the field of bankruptcy prediction by developing the ZETA™ model, a refined multivariate discriminant analysis technique designed to address limitations of the original Altman Z-Score (Altman, 1968). The ZETA model extended the applicability of bankruptcy prediction beyond manufacturing firms, incorporated a more recent and diverse dataset, and included an expanded set of financial ratios—such as return on assets, cumulative profitability, liquidity, and measures of stability and size. Emphasizing improved variable selection and a larger, representative sample, the ZETA model achieved over 90% accuracy in predicting bankruptcy one year before failure, outperforming earlier models and establishing a benchmark for subsequent credit risk assessments across industries.

Lau (1987) introduced a novel multi-state Markov model to capture the dynamic and probabilistic nature of financial distress, moving beyond the traditional binary classification of firms as failed or non-failed. Lau proposed a five-state framework—healthy, bankruptcy, default, private workout, and merger—allowing firms to transition between states over time. By estimating transition probabilities, the model provided a more realistic and nuanced understanding of corporate trajectories. This framework emphasized the temporal evolution of financial distress and offered decision-makers a more flexible tool for risk management and strategic inter-

vention.

Ohlson (1980) made a pivotal methodological contribution by introducing a logistic regression model for bankruptcy prediction, departing from the restrictive assumptions underlying discriminant analysis—such as multivariate normality and equal covariance matrices. His logit model estimated the probability of bankruptcy using financial indicators like firm size, leverage, performance, and liquidity. Applied to a large sample of over 2,000 firms, the model yielded high classification accuracy and introduced probabilistic outputs that proved useful in risk-sensitive environments such as banking and investment.

Building on this foundation, recent literature has increasingly turned to machine learning (ML) methods for bankruptcy prediction due to their ability to model non-linear relationships, handle high-dimensional datasets, and reduce human bias in variable selection. For instance, Lessmann et al. (2015) benchmarked a wide range of ML algorithms and found that ensemble methods such as Random Forests and Gradient Boosting Machines significantly outperformed traditional models in credit scoring tasks. Similarly, Zhou et al. (2021) and Kim et al. (2022) demonstrated the efficacy of XGBoost in predicting corporate distress, particularly when paired with techniques for variable importance and model interpretability.

While traditional statistical models offer transparency and ease of interpretation, modern ML approaches provide superior predictive power, especially when the data exhibit complex patterns or interactions. As the literature increasingly shifts toward hybrid approaches that combine the interpretability of classical models with the accuracy of machine learning, this study contributes by empirically evaluating both classes of models—regularized logistic regressions (Lasso and Ridge) and tree-based ensemble methods (Random Forest and XGBoost)—using real-world financial data.

3 Data, Model, and Methodology

The data for this study came from Theodossiou, Kahya, Saidi, and Philippatos (1996), "Financial Distress and Corporate Acquisitions: Further Empirical Evidence" (Journal of Business Finance and Accounting). According to their research, a company is considered financially distressed if it meets at least one of the following criteria: (1) real debt default, (2) management conversations with creditors to renegotiate terms of debit instruments, or (3) trouble meeting the payment requirements of debt contracts. All data are taken from Compustat and only cover corporations listed on the NYSE or AMEX from 1981 to 1989.

The sample includes 181 manufacturing enterprises, 86 of which were designated as financially distressed at some point throughout the period. Data for distressed enterprises are collected typically one year before they show the first signs of distress. Firm-specific variables employed include size metrics and ratios that reflect liquidity, debt, managerial efficiency, and profitability.

Liquidity refers to a firm's ability to pay its short-term debts using its short-term assets. Managerial efficiency can be gauged using ratios that reflect how effectively the management uses the company's resources. Profitability ratios assess the firm's ability to generate earnings relative to its costs and resources.

We employ a range of statistical and machine learning models to predict corporate bankruptcy using a set of financial distress indicators derived from firm-level accounting ratios. The baseline analysis begins with classical generalized linear model—Logistic (logit) regression—to estimate the likelihood of financial distress as a function of financial ratios such as leverage, profitability, liquidity, and asset composition. These models provide interpretable marginal effects and serve as benchmarks for assessing nonlinear approaches. To address potential multicollinearity and identify the most predictive variables, we also estimate penalized logistic models using Lasso (L1 regularization) and Ridge (L2 regularization) regressions. Hyperparameters are tuned via 10-fold cross-validation, and all predictors are standardized to ensure the penalties are appropriately scaled.

To account for nonlinear relationships and complex interactions between variables, we implement Random Forest and Extreme Gradient Boosting (XGBoost) classifiers. The Random Forest model uses ensemble decision trees trained on bootstrapped samples with randomized feature selection, yielding robust variable importance rankings. XGBoost, a gradient boosting algorithm, sequentially fits decision trees by minimizing the classification loss with regularization to prevent overfitting. We use 5-fold cross-validation to evaluate AUC (Area Under the ROC Curve) across boosting rounds and apply early stopping to prevent performance degradation. Performance metrics such as accuracy, sensitivity, specificity, and AUC are compared across models using out-of-sample predictions to assess generalizability.

Having established the methodological framework, we now present empirical results from each model and evaluate their predictive power in identifying financially distressed firms.

4 Results and Interpretation

This section reports the empirical results from a range of classification models developed to predict corporate bankruptcy using firm-level financial indicators. We begin by estimating interpretable benchmark models—logit regressions—followed by regularized logistic regression using Lasso and Ridge to improve predictive power and handle multicollinearity. Finally, we implement non-linear tree-based models, namely Random Forest and Extreme Gradient Boosting (XGBoost), to capture complex patterns and interactions. Model performance is compared using metrics such as AUC, accuracy, sensitivity, and specificity.

4.1 Logistic Regression Model

Table 1 displays the output of the logistic regression model. Among the predictors, Employee Growth Rate emerges as the most significant, with a negative coefficient (–7.32, p = 0.0015), indicating that firms expanding their workforce face a substantially lower risk of bankruptcy. Debt-to-Assets is positively associated with bankruptcy (p = 0.062), suggesting that higher leverage increases financial vulnerability. Similarly, Operating Income to Assets is negatively associated with distress and marginally significant (p = 0.095). While other financial ratios, such as liquidity and asset composition, are not statistically significant at conventional levels, their signs align with economic intuition. Overall, the logistic regression serves as a robust and interpretable starting point for evaluating financial distress risk.

Table 1: Logit Regression Results

Variable	**Estimate**	**Std. Error**	**Pr(>|z|)**
Intercept	-1.80	2.19	0.41
Debt to Assets	6.84	3.67	0.06$^{\cdot}$
Employee Growth Rate	-7.32	2.30	0.00**
Operating Income to Assets	-13.40	8.03	0.10$^{\cdot}$
Inventory to Sales	8.10	5.64	0.15
Log of Sales	0.21	0.90	0.82
Log of Assets	-0.45	0.89	0.62
Net Working Cap to Assets	-5.13	3.39	0.13
Current Assets to Current Liab	-0.07	0.74	0.93
Quick Assets to Current Liab	1.19	1.00	0.23
EBIT to Assets	9.15	7.33	0.21
Retained Earnings to Assets	-1.08	1.57	0.49
Fixed Assets to Assets	-4.22	3.39	0.21

Note: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

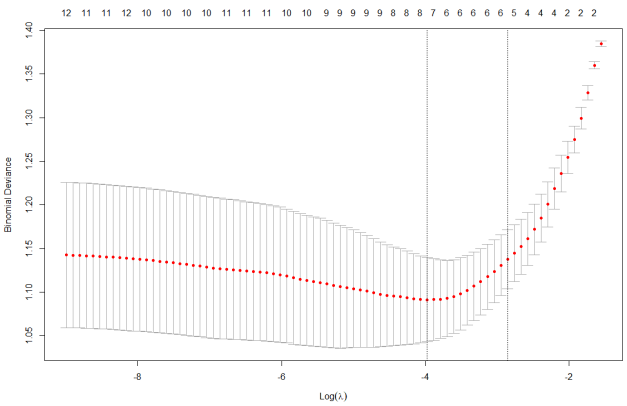

Figure 1: Lasso CV Plot

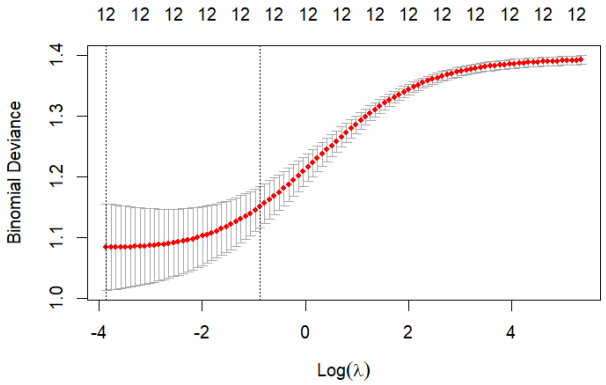

Figure 2: Ridge CV Plot

4.2 Lasso and Ridge Regression

Table 2: Lasso vs. Ridge Coefficients at Optimal Lambda

Variable	Lasso	Ridge
Intercept	-1.67	-0.68
Debt to Assets	4.86	3.62
Employee Growth Rate	-5.06	-5.21
Operating Income to Assets	-4.28	-3.82
Inventory to Sales	3.09	4.24
Log of Sales	-0.14	-0.12
Log of Assets	0.00	-0.05
Net Working Cap to Assets	-0.20	-1.30
Current Assets to Current Liab	0.00	-0.06
Quick Assets to Current Liab	0.25	0.38
EBIT to Assets	0.00	-0.05
Retained Earnings to Assets	-0.76	-1.38
Fixed Assets to Assets	-1.64	-1.52

To improve predictive performance and address multicollinearity among financial ratios, we implement regularized logistic regression models using Lasso (L1 penalty) and Ridge (L2 penalty). Both models introduce a regularization parameter, λ, that penalizes large coefficient estimates, thereby reducing model variance and enhancing generalization. The optimal value of λ is selected through 10-fold cross-validation, minimizing binomial deviance as shown in Figures 1 and 2.

Figure 1 displays the Lasso cross-validation plot, where binomial deviance reaches its minimum around $\log(\lambda) \approx$ -4, indicating an optimal balance between model complexity and predictive accuracy. Notably, Lasso encourages sparsity: it shrinks many coefficients toward zero, effectively performing variable selection. As shown in Table 2, Lasso excludes several variables entirely (e.g., Log of Assets, Current Assets to Liabilities), focusing on a smaller set of predictive features. The most prominent among these are Employee Growth Rate, Debt to Assets, and Operating Income to Assets, all of which exhibit relatively large absolute coefficients.

In contrast, Figure 2 shows the Ridge model's deviance curve, which flattens gradually as λ increases. Unlike Lasso, Ridge does not eliminate variables but instead shrinks all coefficients continuously. This is evident in Table 2, where the

Ridge model retains all predictors but with more moderated estimates. While the magnitude of coefficients is generally smaller than in the Lasso model, Ridge is particularly effective in stabilizing estimates when predictors are highly correlated—a common occurrence with financial ratios.

Lasso's sparsity makes it valuable for interpretation and feature selection, while Ridge offers stability in the presence of multicollinearity. These regularized approaches complement the baseline logit model by improving generalizability and highlighting the most informative financial indicators in predicting corporate distress.

4.3 Random Forest

To capture nonlinear interactions and improve classification performance, we estimate a Random Forest model using 500 decision trees with 3 variables randomly selected at each split. The model achieves an Out-of-Bag (OOB) error rate of 30.17%, indicating reasonably good generalization performance. As seen in Table 3, the confusion matrix shows balanced classification accuracy across both classes, with a class error of ∼30% for both healthy and distressed firms. This reflects the model's ability to distinguish between bankrupt and solvent firms with moderate success, without heavily favoring one class over the other.

Table 3: Random Forest Model Summary

Model Type	Classification
Number of Trees	500
Variables per Split	3
OOB Error Rate	30.17%

Actual / Predicted	**0**	**1**	**Class Error**
0 (Healthy)	65	28	0.301
1 (Distressed)	26	60	0.302

Note: Out-of-Bag (OOB) estimate is based on predictions for unseen observations during training.

Variable importance rankings, displayed in Figure 3, are based on two metrics: Mean Decrease in Accuracy and Mean Decrease in Gini Impurity. Across both metrics, the most influential variables are Operating Income to Assets, Debt to Assets, and EBIT to Assets, reinforcing the relevance of profitability and leverage in predicting financial distress. Notably, these results are consistent with findings from earlier models, suggesting that these indicators carry predictive strength even in more flexible, non-parametric frameworks. In contrast, variables such as Net Working Capital to Assets, Quick Ratio, and Inventory to Sales rank lower, indicating limited marginal contribution once interactions and nonlinearities are captured.

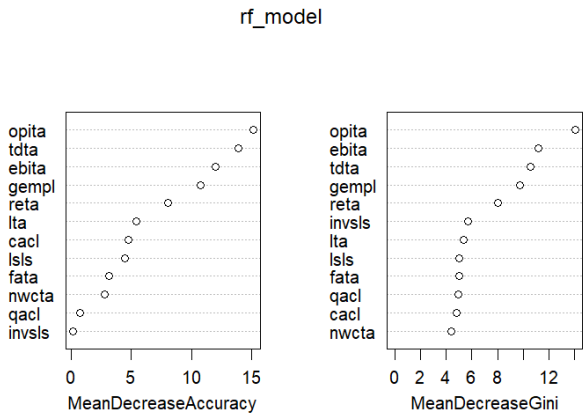

Figure 3: Random Forest Model

4.4 Extreme Gradient Boosting

To further enhance predictive accuracy, we implement an XGBoost classifier, a powerful ensemble learning method that builds decision trees sequentially using gradient-based optimization. The model is tuned through 5-fold cross-validation, and its performance is evaluated using the Area Under the Curve (AUC) metric across 100 boosting iterations. As shown in Table 4, the training AUC rises quickly and reaches near-perfect levels by iteration 9, indicating the model fits the training data extremely well. More importantly, the test AUC peaks at iteration 19 with a value of 0.805, which is the highest out-of-sample performance observed across all models.

Table 4: XGBoost 5-Fold Cross-Validation AUC Scores

Iter	**Train AUC Mean**	**Train AUC Std**	**Test AUC Mean**	**Test AUC Std**
1	0.952	0.005	0.742	0.022
5	0.997	0.002	0.773	0.026
8	1.000	0.000	0.794	0.025
9	1.000	0.000	0.797	0.020
19	**1.000**	**0.000**	**0.805**	**0.023**
20	1.000	0.000	0.799	0.023
30	1.000	0.000	0.797	0.019
50	1.000	0.000	0.796	0.022
100	1.000	0.000	0.793	0.018

Note: Train AUC stabilizes near 1.0 early on; Test AUC peaks at iteration 19 with 0.805, then flattens.

This trend is clearly visualized in Figure 4, which plots the evolution of AUC over boosting rounds. While the train AUC flattens at 1.0 early on, the test AUC improves steadily until iteration 19, after which it plateaus and shows signs of overfitting. This behavior underscores the importance of early stopping, as additional trees beyond iteration 19 do not contribute to generalization and may reduce performance. Compared to other models, XGBoost demonstrates superior test AUC while maintaining low variance across folds, making it the strongest performer in terms of predictive accuracy.

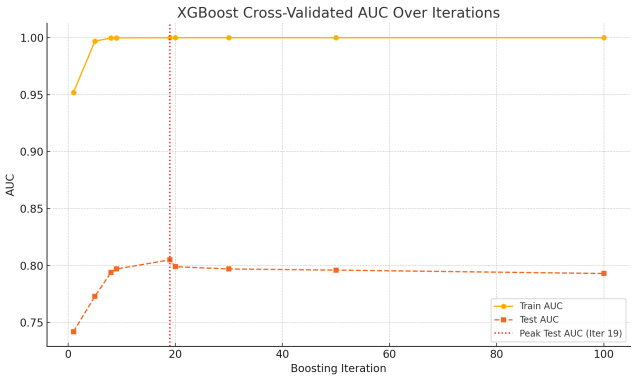

Figure 4: XGBoost AUC Plot

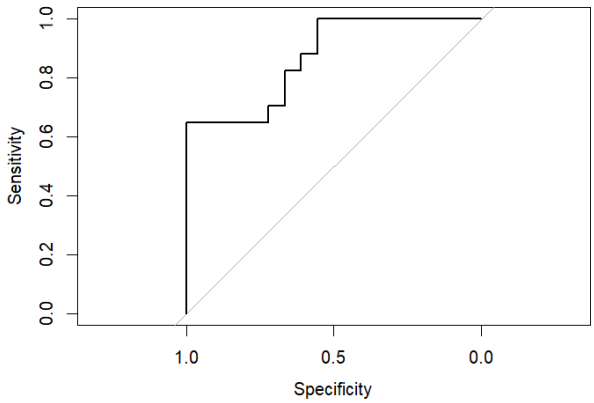

Figure 5: ROC Curve

4.5 Model Evaluation and Final Assessment

Table 5: Confusion Matrix and Classification Statistics

Actual \ Predicted	0	1
0 (Healthy)	12	3
1 (Distressed)	6	14

Metric	Value
Accuracy	0.7429
95% CI	(0.5674, 0.8751)
No Information Rate	0.5143
P-Value [Acc > NIR]	0.004919
Kappa	0.4878
McNemar's Test P-Value	0.504985
Sensitivity	0.6667
Specificity	0.8235
Positive Predictive Value (PPV)	0.8000
Negative Predictive Value (NPV)	0.7000
Prevalence	0.5143
Detection Rate	0.3429
Detection Prevalence	0.4286
Balanced Accuracy	0.7451

Note: The positive class is defined as 0 (non-bankrupt).

To assess the predictive validity of the best-performing model on unseen data, we evaluate its classification performance using a confusion matrix and diagnostic metrics, summarized in Table 5. The model achieves an overall accuracy of 74.3%, with a balanced accuracy of 74.5%, suggesting consistent performance across both classes. The specificity (true negative rate) is 82.4%, indicating that the model effectively avoids false alarms for healthy firms, while the sensitivity (true positive rate) is 66.7%, reflecting its ability to capture distressed firms with reasonable accuracy.

The Kappa statistic of 0.49 indicates moderate agreement beyond chance, and the p-value associated with surpassing the No Information Rate is statistically significant ($p = 0.0049$), confirming that the model adds predictive value over a naive classifier. As shown in Figure 5, the ROC curve reflects strong discriminative ability, with an Area Under the Curve (AUC) of 0.8693, demonstrating that the model can successfully distinguish between bankrupt and non-bankrupt firms across a range of thresholds. These results support the model's practical utility in financial risk screening applications, offering a reliable tool for early identification of firm-level financial distress.

5 Conclusion

This study evaluated the predictive power of various statistical and machine learning models in identifying corporate bankruptcy using firm-level financial indicators. Beginning with interpretable baseline models such as logistic regression, and progressively incorporating regularization (Lasso, Ridge) and tree-based ensemble methods (Random Forest, XGBoost), we examined both model interpretability and predictive performance across a range of evaluation metrics.

Across all models, XGBoost emerged as the most effective classifier, achieving the highest test AUC of 0.805 during cross-validation and a final out-of-sample AUC of 0.8693, outperforming both linear and non-linear counterparts. It demonstrated strong overall accuracy (74.3%), balanced class-wise performance, and the ability to capture complex interactions within financial data. Random Forest also performed competitively, offering robustness and meaningful variable importance rankings, though with a slightly higher error rate.

Among the linear models, Lasso regression proved particularly valuable for variable selection, consistently highlighting key financial ratios such as Employee Growth Rate, Operating Income to Assets, and Debt to Assets—factors which also featured prominently in tree-based models. Ridge regression provided stability in the presence of multicollinearity, but its performance gains were more incremental. While the tradi-

tional logistic regression model was highly interpretable and economically intuitive, it underperformed compared to regularized and non-linear models in terms of classification metrics.

In summary, while linear models like Lasso are suitable for interpretable early-warning systems, XGBoost is best suited for high-stakes prediction tasks requiring greater accuracy and the ability to capture non-linear patterns in financial data. These findings underscore the utility of ensemble learning in financial risk analytics and support the adoption of XGBoost as a practical and theoretically sound tool for bankruptcy prediction.

References

1. Altman, Edward I., Robert G. Haldeman, and P. Narayanan. *ZETA Analysis: A New Model to Identify Bankruptcy Risk of Corporations*. Journal of Banking and Finance, vol. 1, no. 1, 1977, pp. 29–54.

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

3. Breiman, L. (2001). *Random forests*. Machine Learning, 45(1), 5–32.

4. Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. In Proceedings of the 22nd ACM SIGKDD Conference, pp. 785–794.

5. Hoerl, A. E., & Kennard, R. W. (1970). *Ridge regression: Biased estimation for nonorthogonal problems*. Technometrics, 12(1), 55–67.

6. Huang, Z., Chen, H., Hsu, C.-J., Chen, W.-H., & Wu, S. (2004). *Credit rating analysis with SVM and neural networks: A comparative study*. Decision Support Systems, 37(4), 543–558.

7. Jose Manuel Pereira, Mario Basto, Amelia Ferreira da Silva. *The Logistic Lasso and Ridge Regression in Predicting Corporate Failure*. Procedia Economics and Finance, 39, 2016, 634–641.

8. Kim, Hyeongwoo, Youngsoo Kim, and Daejin Kim. *Interpretable ML for Bankruptcy Prediction: A Comparative Analysis Using SHAP*. Expert Systems with Applications, vol. 198, 2022.

9. Lau, Lawrence J. *A Five-State Financial Distress Model*. Journal of Banking & Finance, vol. 11, no. 1, 1987, pp. 1–20.

10. Lessmann, Stefan, et al. *Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring*. European Journal of Operational Research, vol. 247, no. 1, 2015, pp. 124–136.

11. Ohlson, James A. *Financial Ratios and the Probabilistic Prediction of Bankruptcy*. Journal of Accounting Research, 18(1), 1980, pp. 109–131.

12. Rabaca, V., & Silva, J. *Logit Ridge and Lasso in Predicting Business Failure*. Global Journal of Accounting and Economic Research, 3(1), 2023, 25–38.

13. Theodossiou, Panagiotis, Yahya S. Kahya, Ali Saidi, and George C. Philippatos. *Financial Distress and Corporate Acquisitions: Further Empirical Evidence*. Journal of Business Finance & Accounting, 23(5–6), 1996, 699–719.

14. Tibshirani, R. (1996). *Regression shrinkage and selection via the lasso*. Journal of the Royal Statistical Society: Series B, 58(1), 267–288.

15. Tian, S., Yu, Y., & Guo, H. *Variable Selection and Corporate Bankruptcy Forecasts*. Journal of Banking & Finance, 52, 2015, 89–100.

16. Zhou, Lin, Yuxing Liu, and Jin Zhang. *Bankruptcy Prediction Using Machine Learning Algorithms: A Comprehensive Empirical Study*. Computational Economics, 58, 2021, 633–655.