Proceedings of ICFTBA 2026 Symposium: Driving Organizational Strategic Change for Financial Gains
DOI: 10.54254/2754-1169/2026.BH38229

Explainable Machine Learning for Corporate Bankruptcy Prediction: Evidence from Accounting Information

Ziqi Zhao

Honors College, Tianjin Foreign Studies University, Tianjin, China
18622126150@163.com

Abstract. Bankruptcy prediction sits at the core of credit risk management, yet the ML models now used for it are seldom examined for interpretability. This paper benchmarks five classifiers — Logistic Regression, XGBoost, Support Vector Machine, K-Nearest Neighbors, and a Neural Network — on 78,682 firm-year records of U.S. public companies from 1999 to 2018. Boruta selection screens 36 raw items and accounting ratios down to the informative ones, and class imbalance is tackled with cost-sensitive learning plus a Youden's J threshold. XGBoost attains the highest test AUC (0.8343), trailed by ANN (0.8255) and KNN (0.8149). Retained earnings ratio, market-to-assets, and EBIT emerge from SHAP as the strongest predictors. Across pre-crisis (1999–2006), crisis (2007–2009), and post-crisis (2010–2011) windows, the relative importance of these indicators stays largely unchanged, with retained earnings leading in every period. By pairing strong predictive accuracy with explainable evidence, the study narrows the gap between black-box performance and the transparency expected by lenders and regulators. The resulting framework acts as an auditable early-warning tool that practitioners can both trust and interrogate.

Keywords: corporate bankruptcy prediction, accounting information, explainable machine learning, SHAP, financial crisis

1. Introduction

Bankruptcy is costly — for creditors and shareholders, for employees, and for the wider economy. A Chapter 7 or Chapter 11 filing in the United States rarely stays contained: job losses and tighter credit conditions tend to follow. Because stakeholders who can see distress coming are able to act before the damage becomes irreversible, predicting financial failure has remained a central question in accounting and finance for decades [1-3].

The field's foundations are usually traced to Beaver's single-ratio analysis [1] and to Altman's multivariate Z-score [2]; Ohlson later brought logistic regression into the picture, showing that a probability conveys more than a yes-or-no label [3]. These tools still carry weight — Altman's Z-score, five decades on, is still applied in practice. Their weakness is the linearity they assume: real financial distress turns on interactions among accounting variables that a linear score cannot represent.

Over the last ten years machine learning has moved into this literature in force. Ensembles of trees — XGBoost [4] above all — have repeatedly beaten conventional statistical models on

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2754-1169/2026.BH38229

accuracy [5, 6]. Working with French firms, Nguyen et al. [5] compared XGBoost, LightGBM, and Random Forest and found the ensembles ahead of logistic regression at every horizon from one to five years. Yune and Back [6] reported that gradient boosting holds up well even when the classes are severely imbalanced, provided the sampling is handled carefully.

Accuracy by itself, though, is not what regulators, auditors, and creditors ask for. They want to know why a firm has been marked high-risk, not just that the model is usually right. SHapley Additive exPlanations (SHAP) [7] has become the standard way of opening up such black boxes, because it supplies both a global importance ranking and an explanation for each individual case. Finance has taken it up accordingly — to read credit-scoring models [8], to study corporate default [6], and to classify financial distress [9] — closing part of the distance between model output and the trust stakeholders place in it.

Four things set this paper apart. It compares five classifiers that come from genuinely different families — linear (Logistic Regression), boosted trees (XGBoost), kernel methods (SVM), instance-based learning (KNN), and a neural network (ANN) — on a large sample of U.S. public firms. It uses Boruta selection [10] to sort the genuinely informative predictors from 36 candidates that mix raw accounting figures with constructed ratios. It reads SHAP at two levels at once, global and temporal, tracking how the importance of the accounting signals shifts across the pre-crisis, crisis, and post-crisis years. And it treats the imbalance problem explicitly, through cost-sensitive learning and a Youden's J threshold rather than a default cutoff.

2. Dataset and preprocessing

2.1. Dataset description

The empirical work draws on the American Companies Bankruptcy Prediction Dataset available on Kaggle [11], which records accounting data for U.S. companies traded on NYSE and NASDAQ between 1999 and 2018. It holds 78,682 firm-year observations drawn from 8,971 distinct firms. A firm-year is marked 'failed' when the firm entered Chapter 7 or Chapter 11 in the fiscal year after the observation; otherwise it is 'alive'. Eighteen raw items are provided — among them current assets, cost of goods sold, depreciation and amortization, EBITDA, inventory, net income, total receivables, market value, net sales, total assets, longterm debt, EBIT, gross profit, current liabilities, retained earnings, revenue, total liabilities, and operating expenses — each in millions of USD. The outcome is heavily skewed: only 5,220 of the 78,682 observations are failures, a rate of 6.63%.

2.2. Descriptive statistics

Firms in the sample vary enormously, as one would expect given the mix of industries and size classes involved. Total assets (X10) run from close to zero up to several hundred billion dollars, so small caps and large caps are both represented. Net income (X6) averages 201.61 against a median of only 7.02 — a distribution in which a handful of firms accounts for most of the profit. Market value (X8) and total revenue (X16) are likewise right-skewed, a common feature of cross-industry corporate data. On top of the 18 raw items, 18 ratios were computed to cover the usual dimensions: liquidity (current ratio, quick ratio), leverage (debt-to-assets, debt-to-equity, long-term debt ratio), profitability (ROA, ROE, gross, operating, net, and EBITDA margins), efficiency (asset, inventory, and receivables turnover, operating expense ratio), and valuation (market-to-assets, working capital ratio, retained earnings ratio). Raw items and ratios together give a candidate pool of 36 features.

DOI: 10.54254/2754-1169/2026.BH38229

2.3. Preprocessing pipeline

A time-based split produced the three subsets. Training took the 1999–2011 observations (55,927 samples, 7.94% failures), validation the 2012–2014 observations (10,473 samples, 4.69%), and testing the 2015– 2018 observations (12,282 samples, 2.34%), the last of these held out entirely. Splitting by year rather than at random mirrors how such a model would actually be deployed: fitted on the past, scored on periods not yet seen. Features were z-score standardized with a StandardScaler fit on the training set alone, so no information leaks from validation or test. Boruta selection [10], run with a Random Forest of 200 trees and balanced class weights, was then applied to all 36 candidates; it kept every one, which indicates that the raw items and the derived ratios each contribute predictive information about bankruptcy risk.

3. Methodology

3.1. Classification models

Five classification models spanning distinct learning paradigms were evaluated. Logistic Regression was selected as a linear baseline due to its widespread use in traditional accounting research and its inherent interpretability; an L2-regularized version was tuned over a range of penalty strengths C with balanced class weights to account for the asymmetric loss of misclassifying failed firms. XGBoost (Extreme Gradient Boosting) [4] was included as a state-of-the-art tree-based ensemble method known for strong performance on tabular data; hyperparameters including the number of estimators, maximum tree depth, learning rate, and L1 regularization were optimized via grid search, with scale position weight set to the inverse class ratio. Support Vector Machine employed a linear kernel with probability calibration through sigmoid scaling [12]; the regularization parameter C was tuned with balanced class weights. K-Nearest Neighbors [13] provided an instance-based non-parametric comparison; the number of neighbors and distance metric were optimized, with distance-weighted voting. Finally, a Multi-Layer Perceptron artificial neural network (ANN) with ReLU activation was trained using the Adam optimizer, with hidden layer architectures and L2 penalty strength tuned through grid search with early stopping. All models were implemented using scikit-learn [14] and XGBoost [4] libraries. This selection spans linear, tree-based ensemble, kernel, instance-based, and deep learning paradigms, enabling a systematic assessment of whether model complexity translates into meaningful improvements in bankruptcy prediction.

3.2. Evaluation protocol

Models were evaluated using the area under the receiver operating characteristic curve (AUC) [15], area under the precision-recall curve (AUPRC), accuracy, precision, recall, and F1-score. Given the severe class imbalance in bankruptcy prediction (2.34% failure rate in the test set is a realistic reflection of corporate bankruptcy frequency), accuracy alone is misleading; therefore, AUPRC serves as the primary metric for model selection as it prioritizes correct identification of the minority class. To address class imbalance at the decision threshold level, Youden's J statistic (J = sensitivity + specificity - 1) was used to derive an optimal classification threshold for each model, balancing false positive and false negative rates without manual threshold selection. Hyperparameter tuning was conducted via GridSearchCV on the training set with the validation set (2012–2014) serving as the evaluation fold through a PredefinedSplit, ensuring that the test set (2015–2018) remained entirely unseen during model development.

DOI: 10.54254/2754-1169/2026.BH38229

4. Experimental results

4.1. Model performance comparison

Test-set performance for the five models appears in Table 1. XGBoost comes out on top with an AUC of 0.8343; ANN (0.8255), KNN (0.8149), Logistic Regression (0.7694), and SVM (0.7622) follow in that order. The 0.0721 spread between XGBoost and SVM is wide enough to show that the choice of model genuinely matters here [16]. XGBoost also returns the best AUPRC (0.2322), which is the more telling figure when the positive class is this rare. Its edge reflects two things: it can pick up nonlinear interactions among the financial indicators, and its regularization reins in overfitting on the imbalanced training data. The ANN is not far behind (AUC 0.8255), which says that a neural network can extract useful structure from accounting data without heavy feature engineering. KNN, with distance-weighted voting, reaches 0.8149 — consistent with the idea that firms of similar financial profile tend to meet similar fates. The linear baseline fares worst on AUPRC (0.1179), a sign that its linearity assumption is too rigid for the shape of financial distress.

Table 1. Performance comparison of five classification models on the test set (2015–2018). CV AUC computed on validation set (2012–2014). Optimal threshold determined by Youden's J statistic

Model	CV AUC	Test AUC	AUPRC	Opt. Thresh.	Accuracy	Precision	Recall
XGBoost	0.9995	0.8343	0.2322	0.1769	0.7031	0.0614	0.8188
ANN	0.8011	0.8255	0.1473	0.0858	0.7187	0.0619	0.7805
KNN	1.0000	0.8149	0.1507	0.0902	0.7483	0.0662	0.7456
LogisticRegression	0.7580	0.7694	0.1179	0.5547	0.7995	0.0700	0.6167
SVM	0.7472	0.7622	0.1172	0.0827	0.8375	0.0797	0.5645

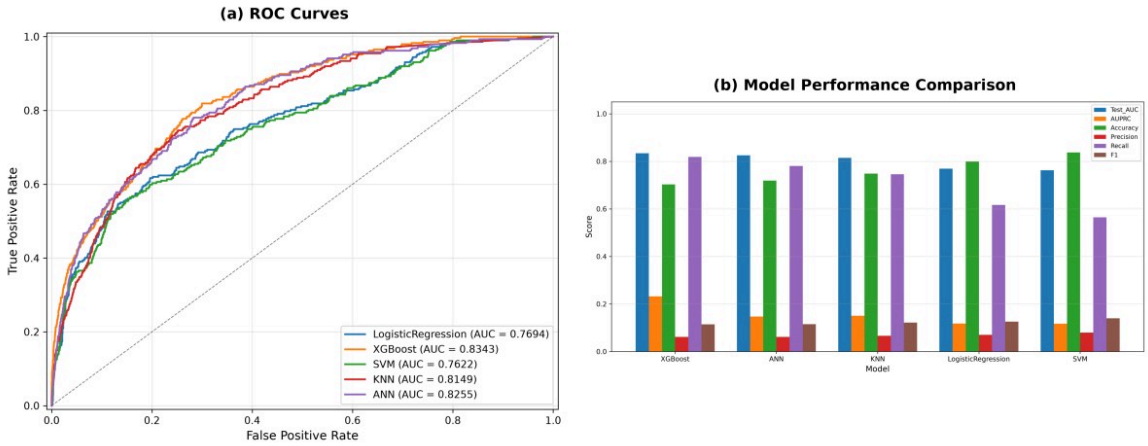

Figure 1. (a) ROC curves for all five models on the test set. (b) Performance comparison across six evaluation metrics

4.2. Analysis of the best model

At its Youden's J threshold of 0.1769, XGBoost reaches a recall of 0.8188, catching 81.9% of the firms that actually failed — but its precision is only 0.0614. That tradeoff is built into the problem: failures make up just 2.34% of the test set, so any high-recall model, however well calibrated, will

DOI: 10.54254/2754-1169/2026.BH38229

raise a large number of false alarms. Which way to lean depends on the use. A system screening a large portfolio of borrowers can tolerate false alarms and should favor recall; a decision on whether to extend credit, where certainty matters more, can instead use a higher cutoff to lift precision. XGBoost's Brier score of 0.0628 points to calibration that is reasonable given the imbalance. The confusion matrix in Figure 1b shows the pattern directly: most failed firms are caught, and overall accuracy stays acceptable.

5. SHAP-based interpretation

5.1. Motivation for interpretability

Strong as its predictions are, XGBoost is still an ensemble of boosted trees — a black box that does not yield its reasoning on inspection. In risk management a probability score alone is not enough; the people relying on it need to see which accounting indicators are moving the prediction. SHAP [7] meets that need by grounding interpretation in cooperative game theory, breaking any single prediction into additive contributions from the Shapley values. It has been taken up widely in finance — credit scoring [8], corporate default [6, 9], financial distress classification [17] — precisely because it works at two scales, giving both an overall ranking of feature importance and a breakdown for one firm's case.

5.2. Global feature importance

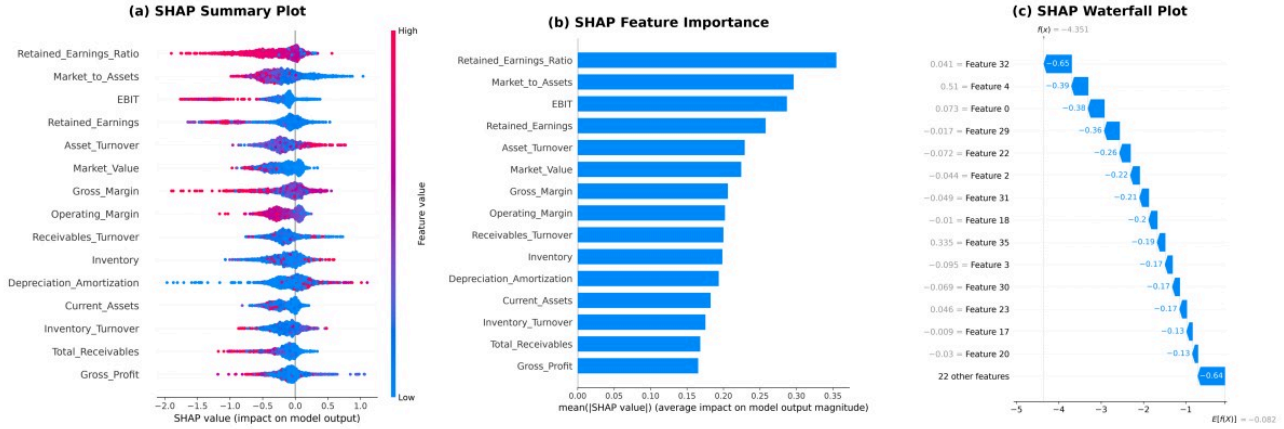

Figure 2. (a) SHAP beeswarm summary plot. (b) Global feature importance (mean |SHAP|). (c) Waterfall plot for a single prediction

TreeExplainer was applied to the fitted XGBoost model on 2,000 test observations. The resulting importance ranking (Figure 2b) places five features at the top: Retained_Earnings_Ratio (0.355), Market_to_Assets (0.296), EBIT (0.287), Retained_Earnings (0.258), and Asset_Turnover (0.229). The leading feature is the retained-earnings ratio (X15 / total assets). This fits accounting intuition: earnings that have been retained, or eaten away, over the firm's whole history are a slow-moving signal, and a shrinking stock of them marks a firm that has been unprofitable for a long while — a condition that often precedes failure. Market-to-assets (X8 / X10) is second; because it sets the market's valuation against book value, a low reading implies the market has already discounted the firm's weakness. EBIT takes third, standing for operating profitability before capital structure and taxes enter. Figure 2a shows the direction of these effects: firms with high retained-earnings and

DOI: 10.54254/2754-1169/2026.BH38229

market-to-assets values sit on the alive side (negative SHAP), while low values push toward failure (positive SHAP) — the expected economic signs.

5.3. Feature-level insights

Looking more closely at the leading features, each carries a clear economic reading. The retained-earnings ratio (importance 0.355) separates firms that have been consistently profitable from those running down their accumulated earnings. A ratio at or below zero deserves attention, and it is exactly the kind of signal a simple metric such as ROA can miss: ROA reports only the current year, whereas retained earnings carry the firm's entire profit history. Market-to-assets (0.296) behaves much like Tobin's Q, folding in growth expectations and intangible value that book figures leave out; that it predicts so well suggests market participants hold forward-looking information that the statements do not. EBIT (0.287) stands for core operating profitability, and asset turnover (0.229) for how efficiently assets are turned into sales — a dimension that is largely independent of both profitability and leverage.

5.4. Crisis-period stability analysis

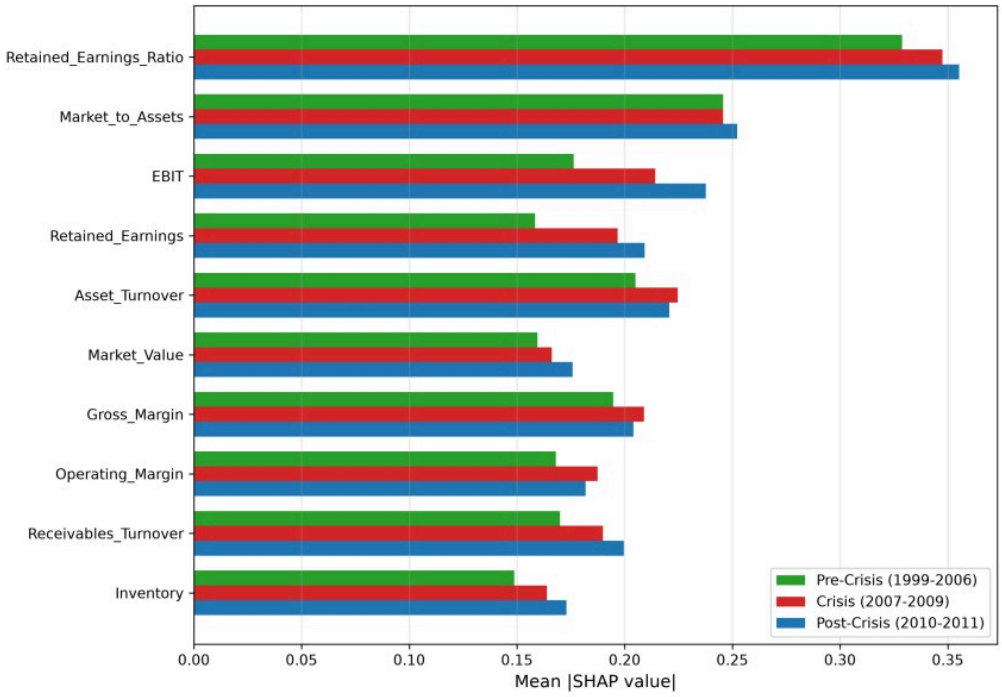

Figure 3. Comparison of SHAP feature importance across pre-crisis (1999–2006), crisis (2007–2009), and post-crisis (2010–2011) periods

Whether these indicators keep their predictive value as the macro environment changes is a separate question, so SHAP values were recomputed on three sub-periods carved out of the training window: precrisis (1999–2006), crisis (2007–2009), and post-crisis (2010–2011). The retained-earnings ratio sits first in all three, and its importance edges up across them — 0.329 before the crisis, 0.347 during it, 0.355 after — which hints that accumulated profitability becomes a slightly sharper signal after an economic shock. Market-to-assets moves in the same mild upward direction (0.246 → 0.246 → 0.252). EBIT climbs from fourth place before the crisis to third during it and to

DOI: 10.54254/2754-1169/2026.BH38229

second afterward, so operational profitability appears to gain relevance once disruption hits. Gross margin ranks higher in the crisis window specifically, plausibly because cost-structure resilience comes to the fore when revenue falls. Taken together the rankings are strikingly similar: the same five features occupy the top in every period, with only small changes in order. That kind of temporal stability is what makes these accounting-based measures workable for long-horizon bankruptcy prediction — their usefulness does not fade with the business cycle.

5.5. Probability calibration

Figure 4 plots the calibration curves. For XGBoost the curve falls below the diagonal in the middle of the range (0.1–0.4), meaning it overstates bankruptcy probabilities there. Under such strong imbalance this is unsurprising: training with a scale_pos_weight tuned to the minority class tends to inflate the probabilities it assigns to that class. The practical consequence is that a raw XGBoost output is better read as a ranking of relative risk than as a literal probability — something worth keeping in mind wherever regulatory settings demand decisions at a stated probability level [18].

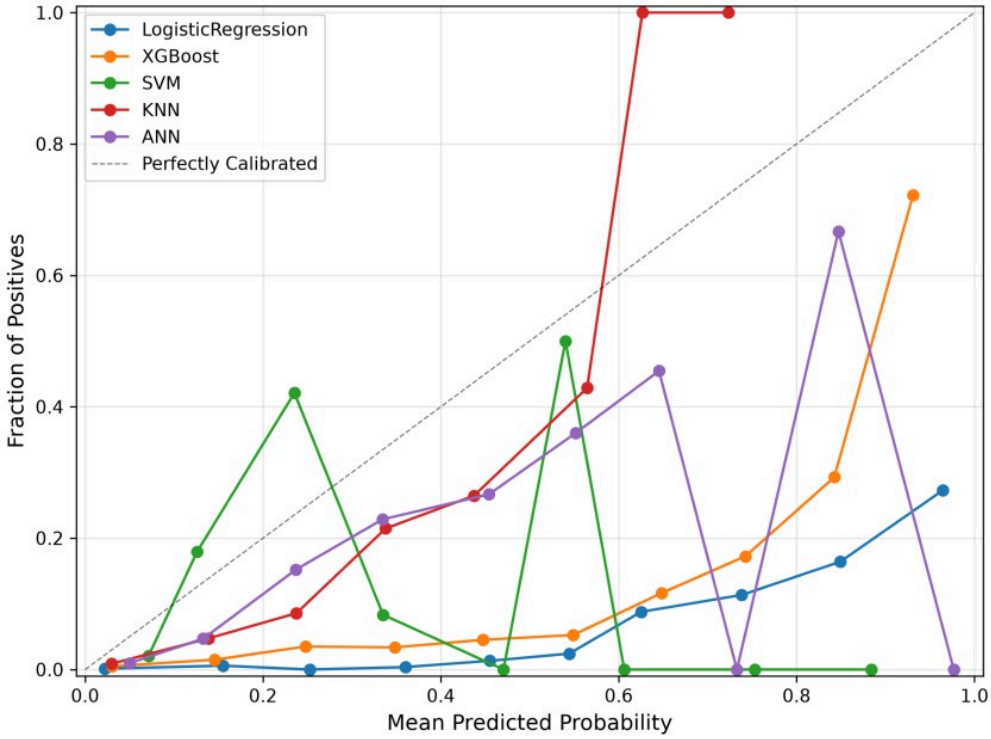

Figure 4. Probability calibration curves for all five models. The diagonal represents perfect calibration

6. Conclusion and future work

Five machine learning classifiers were compared for bankruptcy prediction on U.S. public-company accounting data split by time across 1999–2018. XGBoost gave the best results (AUC 0.8343, AUPRC 0.2322); the linear baseline (Logistic Regression, AUC 0.7694) shows how much the nonlinear models add here. From the SHAP analysis, retained earnings ratio, market-to-assets, and EBIT emerge as the leading predictors, and their ranking holds through the pre-crisis, crisis, and

DOI: 10.54254/2754-1169/2026.BH38229

post-crisis periods — evidence that accounting indicators keep their diagnostic power right across the business cycle [19].

A number of limitations should be noted. The data cover U.S. listed firms only, so the results need not carry over to private companies or to jurisdictions that follow other accounting standards [20]. The test period is so imbalanced (2.34% failures) that high recall comes with low precision; this follows from the real-world frequency of bankruptcy rather than from anything in the models. Boruta kept all 36 features, yet related ratios are correlated to some degree, and reducing that redundancy might make the model more stable. Finally, no macroeconomic variables — interest rates, GDP growth, market indices — were included, even though they plausibly moderate the link between firm-level accounts and bankruptcy risk.

Several extensions suggest themselves. Adding macroeconomic indicators as features could make the model more robust over the cycle. Bringing in textual data — financial disclosures and earnings-call transcripts — would show whether qualitative information adds anything to the quantitative ratios. Testing against more advanced deep learning designs, particularly temporal models that use the panel dimension of the data, might expose dynamic patterns that annual snapshots overlook. And building an interactive SHAP dashboard aimed at credit analysts would help move a research-grade model closer to a tool that can actually be used.

References

[1] Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research, 4*, 71–111. DOI: 10.2307/2490171

[2] Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance, 23*(4), 589–609. DOI: 10.1111/j.1540-6261.1968.tb00843.x

[3] Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research, 18*(1), 109–131. DOI: 10.2307/2490395

[4] Chen, T. & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD* (pp. 785–794). DOI: 10.1145/2939672.2939785

[5] Nguyen, H. H., Viviani, J. L. & Ben Jabeur, S. (2025). Bankruptcy prediction using machine learning and Shapley additive explanations. *Review of Quantitative Finance and Accounting, 65*(1), 107–148. DOI: 10.1007/s11156-023-01192-x

[6] Yune, J. C., & Back, D. H. (2025). eXplainable artificial intelligence applied to corporate bankruptcy prediction with severely imbalanced data. *Journal of Korean Society of Industrial Engineering, 48*(1), 109–128. DOI: 10.11627/jksie.2025.48.1.009

[7] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765–4774).

[8] Alonso-Robisco, A., & Carbo, J. M. (2025). Should we trust the credit decisions provided by machine learning models? *Computational Economics, 66*(5), 4245–4274. DOI: 10.1007/s10614-025-10855-x

[9] Kristanti, F. T. et al. (2024). Advancing financial analytics: Integrating XGBoost, LSTM, and Random Forest algorithms for precision forecasting of corporate financial distress. *Journal of Infrastructure, Policy and Development, 8*(8), 4972. DOI: 10.24294/jipd.v8i8.4972

[10] Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software, 36*(11), 1–13. DOI: 10.18637/jss.v036.i11

[11] Singh, U. (2022). US company bankruptcy prediction dataset. *Kaggle*. https://www.kaggle.com/datasets/utkarshx27/american-companies-bankruptcy-prediction-dataset

[12] Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. DOI: 10.1007/BF00994018

[13] Cover, T. & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21–27. DOI: 10.1109/TIT.1967.1053964

[14] Pedregosa, F. et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

DOI: 10.54254/2754-1169/2026.BH38229

[15] Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145–1159. DOI: 10.1016/S0031-3203(96)00142-2

[16] Gabrielli, G., Melioli, A. & Bertini, F. (2026). Corporate financial distress prediction: A machine learning approach in the era of big data. *Journal of Accounting & Organizational Change, 22*(7), 31–65. DOI: 10.1108/JAOC-052025-0166

[17] Thota, N. et al. (2025). The influence of public listing on bankruptcy prediction in India: An AI-ML approach. *Borsa Istanbul Review, 25*(6), 1463–1475. DOI: 10.1016/j.bir.2025.10.003

[18] Chan, C. P., Tang, F. K. & Yang, J. H. (2026). Financial performance prediction and stability analysis using SHAP-enhanced machine learning models. *Computational Economics*. DOI: 10.1007/s10614-025-11283-7

[19] Bluwstein, K. et al. (2023). Credit growth, the yield curve and financial crisis prediction: Evidence from a machine learning approach. *Journal of International Economics, 145*, 103773. DOI: 10.1016/j.jinteco.2023.103773

[20] Kristanti, F. T., Salim, D. F. & Ihsan, A. F. (2025). Financial distress prediction in emerging markets by using a machine learning approach for financial sustainability. *Discover Sustainability, 6*(1), 1–28. DOI: 10.1007/s43621-025-02199-1