A Two-Stage Machine Learning Approach to Bankruptcy Prediction: From Comprehensive Modeling to Feature Selection for Noise Reduction

Masanobu Matsumaru * and Hideki Katagiri

Posted Date: 30 October 2025

doi: 10.20944/preprints202510.2374.v1

Keywords: corporate bankruptcy prediction; feature selection; ensemble learning; random forest; LightGBM; imbalanced data; Tokyo Stock Exchange

Article

A Two-Stage Machine Learning Approach to Bankruptcy Prediction: From Comprehensive Modeling to Feature Selection for Noise Reduction

Masanobu Matsumaru [1,*] and Hideki Katagiri [2]

[1] Research Institute for Engineering, Kanagawa University, Yokohama 221-8686, Japan

[2] Department of Industrial Engineering and Management, Kanagawa University, Yokohama 221-8686, Japan

* Correspondence: pt120916cz@kanagawa-u.ac.jp or yfa67973@nifty.com

Abstract

Corporate bankruptcy prediction has become increasingly critical amid economic uncertainty. This study proposes a novel two-stage machine learning approach to enhance bankruptcy prediction accuracy, applied to Tokyo Stock Exchange-listed companies. First, models were trained using 173 financial indicators. Second, a wrapper-based feature selection process was employed to reduce dimensionality and eliminate noise, thereby identifying an optimal seven-feature set. Two ensemble learning methods, Random Forest and LightGBM, were used. Random Forest correctly predicted 566 bankruptcies using the reduced feature set (88 more than when using all features) compared with 451 by LightGBM (31 more than when using all features). The study also addresses challenges posed by imbalanced data by employing resampling techniques (SMOTE, SMOTE-ENN, and KMeans). Additionally, the need for industry-specific modeling is recognized by constructing models for the six industry sectors. These findings highlight the importance of feature selection and ensemble learning for improving model generalizability and uncovering industry-specific patterns. This study contributes to the field of bankruptcy prediction by providing a robust framework for accurate and interpretable predictions for both academic research and practical applications. Future work will focus on further enhancing prediction accuracy to identify more potential bankruptcies.

Keywords: corporate bankruptcy prediction; feature selection; ensemble learning; Random Forest ; LightGBM; imbalanced data; Tokyo Stock Exchange

1. Introduction

In today's increasingly uncertain environment, corporate management faces significant challenges due to the heightened risk of bankruptcy arising from deteriorating business performance. Bankruptcies impose substantial losses on stakeholders, including business partners, investors, and financial institutions. Accordingly, developing models that prevent or enable the early detection of bankruptcy has become essential. While traditional research has relied on statistical approaches, recent advances in machine learning have enabled more objective and accurate predictions.

This study builds on earlier work by applying ensemble learning methods—Random Forest and LightGBM—while addressing key challenges such as feature selection, imbalanced data, and industry-specific modeling. In particular, the study integrates resampling techniques with stepwise feature selection, thereby enhancing model generalization, interpretability, and the ability to uncover sector-specific bankruptcy patterns.

Corporate bankruptcy prediction has long attracted scholarly and practical attention. Foundational studies, including Beaver (1966), Altman (1968), Ohlson (1980), and Zmijewski (1984), demonstrated the predictive power of accounting ratios and introduced discriminant and probit/logit models (Nam and Jinn, 2000). Subsequent refinements included hazard models (Shumway, 2001),

 doi:10.20944/preprints202510.2374.v1

industry effects (Chava and Jarrow, 2004), market-based indicators (Hillegeist et al., 2004; Wu, Gaunt, and Gray, 2010), and hybrid approaches integrating accounting, market, and macroeconomic variables (Hernández Tinoco and Wilson, 2013; Ciampi, 2015; Martín and Rojo, 2019). Altman himself and subsequent studies have proposed variations of the Z-score model to extend its applicability to non-manufacturing firms, private companies, and firms in emerging markets (Altman et al., 2017). Lin and McClean (2001) and Varetto (1998) applied data mining and genetic algorithms, while Sun and Shenoy (2007) employed Bayesian networks. Huang et al. (2004) conducted a comparative study using support vector machines and neural networks for credit rating analysis, highlighting the effectiveness of advanced machine learning techniques for financial risk assessment. More recently, Choi and Lee (2018) applied multi-label learning techniques for corporate bankruptcy prediction, demonstrating the potential of advanced machine learning methods to handle multiple bankruptcy indicators simultaneously. These studies provided interpretability but faced limitations in capturing nonlinearities and complex interactions.

Since the 2000s, machine learning has transformed bankruptcy prediction. Ensemble methods, including Breiman's (2001) Random Forest and Dietterich's (2000) ensemble strategies, demonstrated the advantages of combining classifiers. Boosting algorithms such as AdaBoost (Heo and Yang, 2014), Gradient Boosting (Chen and Guestrin, 2016), and LightGBM (Ke et al., 2017) proved highly effective. Hybrid frameworks integrated multiple learners (Wang et al., 2010; Tang and Yan, 2021; Yu and Zhang, 2020), while CatBoost-based models further enhanced accuracy (Ben Jabeur et al., 2021; Mishra and Singh, 2022). Comparative studies confirm that ensemble models consistently outperform single classifiers in both stability and predictive performance (Bandyopadhyay and Lang, 2013; Lahsasna et al., 2018; Lee and Chen, 2020; Kraus and Feuerriegel, 2019; Guillén and Salas, 2021). Stacking and meta-learning frameworks have also been proposed (Tsai and Hsu, 2013; Tang, Zhang, and Chawla, 2009). Moreover, multi-industry frameworks such as Lee and Choi (2013) demonstrate the importance of accounting for sectoral heterogeneity in model design.

A persistent methodological challenge is class imbalance. Bankruptcies are rare compared with solvent cases, and classifiers trained on such datasets often achieve inflated accuracy while failing to capture true defaults. Early work by He and Garcia (2009) and Huang and Ling (2005) highlighted methodological biases, and later studies proposed over- and under-sampling approaches. SMOTE (Chawla et al., 2002), SMOTE-ENN (Batista et al., 2004; Liu and Wu, 2021), and SMOTE-IPF (Sáez et al., 2015) effectively rebalance class distributions. Undersampling approaches such as exploratory undersampling (Liu, Wu, and Zhou, 2009) provided alternatives. Reviews confirm the importance of these methods (Fernandez, García, and Herrera, 2018; Fernández et al., 2018; Sun et al., 2014). Neural network studies further revealed imbalance sensitivity (Buda, Maki, and Mazurowski, 2018). More recent work integrated rebalancing with ensemble learning (Zhao and Wang, 2018; Shetty, Musa, and Brédart, 2022). Recent advances also include hybrid approaches that integrate evolutionary algorithms with domain adaptation (Ansah-Narh et al., 2024).

Feature engineering and dimensionality reduction are also critical. While early studies relied on accounting ratios, more recent work incorporated diverse features such as governance (Ciampi, 2015), market variables, and macroeconomic indicators (Sun et al., 2014). Feature selection reduces noise and overfitting (Guyon and Elisseeff, 2003; Tsai, 2009; Jain and Johnson, 2020). Hybrid strategies (Ko, Kim, and Kang, 2021; Pramodh and Ravi, 2016; Hossari and Rahman, 2022; Kotze and Beukes, 2021; García et al., 2019), explainable AI (Lundberg and Lee, 2017; Wu and Wang, 2022; Zhang and Wang, 2021; Giudici and Hadji-Misheva, 2022; Yeh, Chi, and Lin, 2022), and advances in scikit-learn (Pedregosa et al., 2011) have further improved transparency and reproducibility. In addition, Dikshit and Pradhan (2021) demonstrated the applicability of interpretable models in environmental risk prediction, highlighting the transferability of explainable AI across domains.

Deep learning has expanded the methodological toolkit. Neural networks were introduced by Atiya (2001), Lin, Chiu, and Tsai (2010), and Kim (2011), with more recent contributions focusing on sequential and attention-based models (Li, Sun, and Wu, 2022; Kim, Cho, and Ryu, 2022). Ensemble deep learning reviews further highlight the growing promise of these methods (Ganaie and Hu, 2022).

 doi:10.20944/preprints202510.2374.v1

BERT adaptations extend to textual features (Kim and Yoon, 2023), while hybrid frameworks combine neural networks with market and macroeconomic data (Zhang and Wang, 2021). Comparative studies highlight their growing competitiveness (Iparraguirre-Villanueva and Cabanillas-Carbonell, 2023). Industry- and country-specific determinants remain key: Korean construction (Heo and Yang, 2014), French SMEs (Mselmi, Lahiani, and Hamza, 2017), crisis contexts (Nam and Jinn, 2000; Narvekar and Guha, 2021), dotcom failures (Chandra, Ravi, and Bose, 2009), firm size and volatility in Pakistan (Rashid, Hassan, and Karamat, 2021), policy uncertainty (Fedorova et al., 2022), and early banking distress detection (Ravisankar and Ravi, 2010) further illustrate contextual variation.

Several meta-studies have synthesized these developments. Bellovary, Giacomino, and Akers (2007) reviewed early research, while Ravi Kumar and Ravi (2007), Alaka et al. (2018), Sun et al. (2014), and Dasilas and Rigani (2024) provided systematic reviews. Xu and Ouenniche (2018) applied DEA benchmarking, Radovanovic and Haas (2023) assessed socio-economic costs, and Razzak, Imran, and Xu (2019) reviewed deep learning for credit scoring. Succurro, Arcuri, and Costanzo (2019) introduced robust PCA, and Sánchez-Medina et al. (2024) investigated bankruptcy resolution prediction. Wu, Gaunt, and Gray (2010) compared alternative models, while Yeh and Lien (2009) analyzed credit card default risks in relation to corporate bankruptcy prediction. Du Jardin (2016) further demonstrated the effectiveness of a two-stage classification strategy, motivating the two-stage approach of this study.

Overall, the literature reveals a clear progression from interpretable statistical models to increasingly complex, data-driven, and explainable frameworks. Prediction success depends not only on algorithmic choice but also on handling imbalanced data, selecting informative features, and incorporating industry-specific determinants. Building on this understanding, the present study adopts a two-stage machine learning framework for bankruptcy prediction among Tokyo Stock Exchange–listed firms. In the first stage, comprehensive learning is performed using 173 financial indicators. In the second stage, wrapper-based feature selection is applied to gradually reduce dimensionality, eliminate noise, and arrive at an optimal seven-feature set. This approach enhances both predictive performance and interpretability.

To capture sector-specific heterogeneity, separate models are constructed for six industries—Construction, Real Estate, Services, Retail, Wholesale, and Electrical Equipment—thus uncovering sectoral bankruptcy determinants and patterns. In addition, three resampling techniques—SMOTE, SMOTE-ENN, and k-means clustering—are incorporated to address class imbalance. Empirical results reveal that Random Forest correctly predicts 566 bankruptcies and LightGBM predicts 451, both substantially outperforming models without feature reduction. By simultaneously addressing four key challenges—methodological choices, dataset design, class imbalance, and industry heterogeneity—this study underscores both its novelty and practical relevance.

2. Materials and Methods

2.1. Machine Learning Methods

This study employs Random Forest and LightGBM. Brief explanations of each method follow.

2.1.1. Random Forest

Random Forest is a machine learning framework based on decision-tree algorithms. It uses bagging, which is a type of ensemble learning, to create high-accuracy learners. Bagging trains multiple ‘weak learners’ in parallel using bootstrapping. Bootstrapping extracts data samples with replacement from the original dataset. Since sampling with replacement allows for duplicates, the same data point may be selected multiple times for training a single weak learner. Bagging uses bootstrapping to train multiple weak learners in parallel. Bootstrapping involves resampling with replacement from the original dataset. Each learner performs learning and prediction independently;

 doi:10.20944/preprints202510.2374.v1

for regression tasks, learner predictions are averaged, and for classification tasks, a majority vote determines the final prediction.

2.1.2. LightGBM

LightGBM is also a machine learning framework based on decision tree algorithms. It employs gradient boosting, an ensemble-learning method, to create high-accuracy learners. This approach constructs a 'strong learner' by sequentially combining individual learners. Specifically, it uses the steepest descent method to minimize the error between predicted and actual numbers. XGBoost, a conventional machine learning algorithm that uses gradient boosting, employs a level-wise tree-growth strategy to grow decision trees. This method grows a decision tree by expanding its levels (layers). In contrast, LightGBM employs leaf-wise growth, which grows decision trees by expanding the tree's leaves. This enables LightGBM to achieve faster processing, while maintaining XGBoost's accuracy.

2.2. Evaluation Metrics

Evaluation metrics are quantitative indicators used to assess model performance. Using evaluation metrics helps to determine how accurately the constructed model can predict bankruptcy from the input data. Furthermore, evaluation metrics play a crucial role in comparing model performance and facilitating improvements.

The following evaluation metrics are used in this study. We define true positives (TPs) as instances when bankrupt companies are correctly predicted as bankrupt; false negatives (FNs) as instances when bankrupt companies are incorrectly predicted as non-bankrupt; true negatives (TNs) as instances when non-bankrupt companies are correctly predicted as non-bankrupt; and false positives (FPs) as instances when non-bankrupt companies are incorrectly predicted as bankrupt. Using these TPs, FPs, TNs, and FNs, we calculate the following metrics: precision, accuracy, recall, false positive rate, and false negative rate. Given our focus on bankruptcy, TPs are our primary metric to represent correctly predicted bankrupt companies. Therefore, we consider the model that yielded the highest number of TPs as the best. We compute recall as the evaluation metric for industry-specific bankruptcy prediction, considering different dataset compositions and resampling methods. Since the actual number of bankrupt companies varies by industry, TPs alone cannot determine prediction accuracy. Assessing recall therefore provides a better understanding of bankruptcy prediction accuracy for each industry.

2.3. Numerical Experiment Design

2.3.1. Datasets

We construct models for six industries based on the Tokyo Stock Exchange's sector classifications: construction, real estate, services, retail, wholesale, and electrical equipment. Data were obtained from the financial statements of companies listed on the Tokyo Stock Exchange, using the Nikkei NEEDS database. In this study, we define bankruptcy as a company being delisted owing to civil rehabilitation proceedings or similar events. We use data from 317 companies that filed for bankruptcy between 1991 and 2021.

Table 1 presents three study datasets with different numbers of features. Numbers in parentheses indicate the feature counts. The significant disparity between the numbers of non-bankrupt and bankrupt companies indicates an imbalanced dataset. The first dataset configuration contains data from 52,950 companies and uses only financial indicators as features. The second configuration contains data from 26,674 companies. This dataset augments the short-term financial performance indicators from the first dataset with investment-financing network indicators representing financing diversity and long-term intercompany trust relationships. The third dataset, constructed for benchmarking against the first (financial) and second (investment-financing) datasets, includes data from 26,674 companies. This is derived by stripping the 12 investment-financing

 doi:10.20944/preprints202510.2374.v1

network indicators from the second dataset, leaving 161 purely financial features. Consequently, it differs from the first dataset in terms of size (26,674 vs. 52,950 companies) and from the second in terms of feature count (161 vs. 173 features).

Table 1. Datasets.

Industry	Financial			Investment-financing			Comparison		
	Bankrupt	Non-bankrupt	Total	Bankrupt	Non-bankrupt	Total	Bankrupt	Non-bankrupt	Total
Construction	42	6,280	6,322	25	3,848	3,873	25	3,848	3,873
Real estate	34	3,603	3,637	21	2,089	2,110	21	2,089	2,110
Service	25	11,445	11,470	5	4,480	4,485	5	4,480	4,485
Retail	23	11,531	11,554	10	5,808	5,818	10	5,808	5,818
Electrical equipment	23	9,043	9,066	11	4,722	4,733	11	4,722	4,733
Wholesale	16	10,935	10,951	5	5,650	5,655	5	5,650	5,655
Total	163	52,787	52,950	77	26,597	26,674	77	26,597	26,674

2.3.2. Indicators

We utilize all 161 financial indicators available from the Nikkei NEEDS-Financial QUEST (FQ), a comprehensive economic database service. Following the NEEDS-FQ classification system, these indicators are categorized into seven types: profitability, return on capital, margin-related, productivity, stability, growth, and cash flow indicators (Table 2).

Table 2. Financial indicators and features.

Classification of financial indicators	Number of features	Examples of features
Profitability	47	Profit margin
Return on capital	15	Return on assets
Margin related	10	EBIT [1] margin
Productivity	6	Revenue per employee
Safety	35	Equity ratio
Growth	15	Revenue growth rate (YOY [2])
Cash flow	33	Cash flow to net debt ratio

[1] EBIT: earnings before interest and taxes. [2] YOY: years.

2.3.3. Investment-Financing Network Indicators

We calculate investment-financing network indicators from networks representing corporate investment and financing relations. To construct these networks, we use data from the Nikkei NEEDS-Financial QUEST on major shareholders, corporate shareholdings, and loans. Specifically, we calculate six investment network indicators and six financing network indicators, totaling 12 indicators.

We define each indicator as follows:

- Degree centrality: How connected a given node is to other nodes in a network.
- Betweenness centrality: The frequency with which a given node lies on the shortest path between other nodes, indicating the degree of connection between the node and others.
- Network density: The degree of connection among nodes, expressed as a ratio. The denominator is the number of possible connections, and the numerator is the number of actual connections.
- Authority score: The extent to which other nodes link to a given node, representing how many other nodes are linked to it.
- Hub score: Outgoing edge connections to other nodes.
- PageRank: The importance of a webpage.

 doi:10.20944/preprints202510.2374.v1

2.3.4. Computational Environment

Before running the models with LightGBM and Random Forest, we perform data standardization as a preprocessing step. Next, we conduct a 10-fold cross-validation to optimize the hyperparameters using Optuna, a Python package. The prediction models are then constructed using the optimized parameters. Tables 3 and 4 present the corresponding hardware and software specifications.

Table 3. Hardware and software specifications.

Hardware/Software	Specification
CPU [1]	Core i9-10885H
RAM [2]	32.0 GB[4]
OS [3]	Windows 11 Pro
Programming Language	Python 3.9.7
Machine Learning Package	scikit-learn
Parameter Optimization Package	Optuna

[1] CPU, central processing unit. [2] RAM, random-access memory. [3] OS, operating system. [4] GB, gigabytes.

3. Results

In this two-stage study, the results of the second stage are derived from a dataset of seven features obtained through progressive feature reduction of the 173-feature set used in the first stage. This eliminates noise and prevent overfitting. We use the feature importance attribute to determine which features the models rely on the most for bankruptcy prediction, quantifying each feature's importance and visualizing the results.

3.1. First Stage Results (173 Features)

This section presents the results of the first stage of bankruptcy prediction. In the first stage, we construct bankruptcy prediction models using Random Forest and LightGBM with a dataset containing all 173 available financial indicators.

3.1.1. Random Forest

Table 4 presents the results obtained by dataset configuration using the 173-feature set. The total TP count across the financial, investment financing, and comparison datasets was 478. Among the datasets, the financial dataset, having the largest sample size, achieved the best prediction accuracy, with 284 TPs. Both the investment financing and comparison datasets yielded 97 TPs, indicating no apparent advantage in using investment financing network indicators.

Table 4. Random Forest true positive count by dataset configuration.

Actual	Predicted			
	Total	Financial	Investment-financing	Comparison
951	478	284	97	97

Resampling: For the 317 bankrupt companies across the six industries (Table 5), k-means achieved the highest accuracy with 186 TPs across all three datasets, followed by SMOTE-ENN with 177 TPs, and SMOTE with 115 TPs.

Table 5. Random Forest true positive count by resampling method and dataset configuration.

Resampling	Actual	Predicted			
		Total	Financial	Investment-financing	Comparison
SMOTE	317	115	74	19	22

 doi:10.20944/preprints202510.2374.v1

SMOTE-ENN	317	177	100	40	37
Kmeans	317	186	110	38	38
Total	951	478	284	97	97

Industry-Specific Results: Table 6 presents the industry-specific prediction results by dataset configuration. The highest TP count (75) was recorded in the construction industry when the financial dataset was used. By contrast, the lowest TP count (0) was recorded for the wholesale industry when using the comparison dataset.

Table 6. Random Forest true positive count by industry and dataset configuration.

Industry	Total			Financial			Investment-financing			Comparison		
	Actual	Predicted	Recall	Actual	Predicted	Recall	Actual	Predicted	Recall	Actual	Predicted	Recall
Construction	276	152	55.07	126	75	59.52	75	37	49.33	75	40	53.33
Real estate	228	133	58.33	102	69	67.65	63	31	49.21	63	33	52.38
Service	105	55	52.38	75	50	66.67	15	3	20.0	15	2	13.33
Retail	129	58	44.96	69	39	56.52	30	10	33.3	30	9	30.00
Electrical equipment	135	59	43.70	69	31	44.93	33	15	45.45	33	13	39.39
Wholesale	78	21	26.92	48	20	41.67	15	1	6.67	15	0	0.00
Total	951	478	50.26	317	284	89.59	317	97	30.6	317	97	30.60

Resampling Results by Industry: Table 7 presents the resampling results by industry. Applying k-means resampling to the construction industry data yielded the highest TP count (59), whereas applying SMOTE resampling to the wholesale industry data produced the lowest TP count (3).

Table 7. Random Forest true positive count and recall rate by resampling method and industry.

Industry	Total			SMOTE			SMOTE-ENN			K-means		
	Actual	Predicted	Recall	Actual	Predicted	Recall	Actual	Predicted	Recall	Actual	Predicted	Recall
Construction	276	152	55.07	92	35	39.13	92	57	61.96	92	59	64.13
Real estate	228	133	58.33	76	37	48.68	76	50	65.79	76	46	60.53
Service	105	55	52.38	35	15	42.86	35	18	51.43	35	22	62.86
Retail	129	58	44.96	43	9	20.93	43	24	55.81	43	25	58.14
Electrical equipment	135	59	43.70	45	15	33.33	45	20	44.44	45	24	53.33
Wholesale	78	21	26.92	26	3	11.54	26	8	30.77	26	10	38.46
Total	951	478	50.26	317	115	36.28	317	177	55.84	317	186	58.68

Taken together, the results in Tables 6 and 7 show that applying Random Forest to the data from the construction industry yields the highest total TP count (152), whereas applying this method to the wholesale industry data produces the lowest total TP count (21). Recall was highest in the real estate industry (58.33%) and lowest in the wholesale industry (26.92%).

Feature Importance: To pinpoint which features were significant for the model's predictions, we visualized the feature importance scores assigned by the Random Forest algorithm. Figure 1 illustrates the top seven features (out of 173) for the electrical-equipment industry using the financial dataset with SMOTE-ENN resampling.

 doi:10.20944/preprints202510.2374.v1

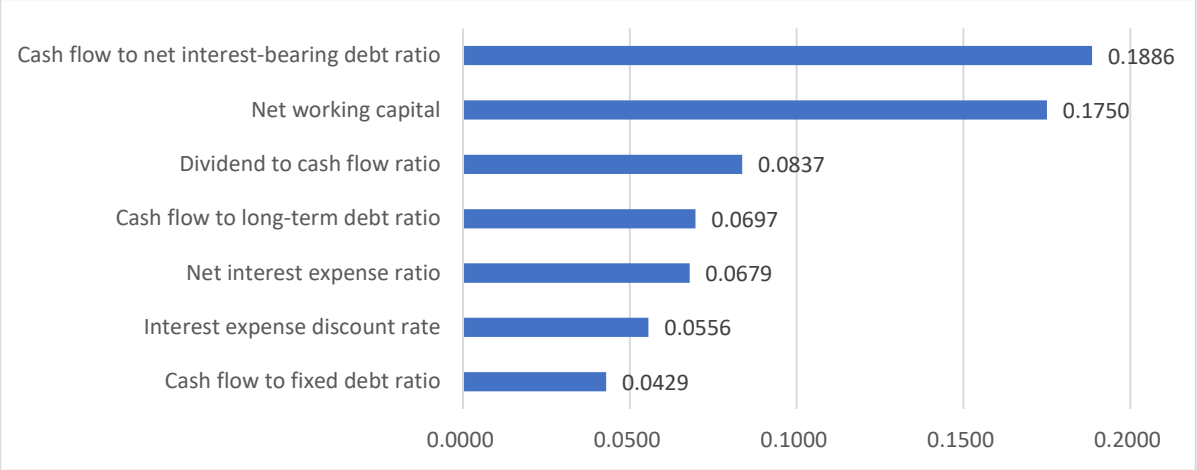


Figure 1. Random Forest feature importance scores for 173 features under SMOTE-ENN resampling using the financial dataset for the electrical equipment industry.

3.1.2. LightGBM

Table 8 presents the results obtained using 173-feature set. Across the financial, investment-financing, and comparison datasets, Random Forest correctly identified 420 bankrupt firms, 58 more than LightGBM. When broken down by dataset, Random Forest achieved its highest TP count on the financial dataset, which had the largest sample size with 264 TPs, and identified 58 bankrupt firms using the investment-financing dataset and 97 using the comparison dataset. Notably, we found no significant differences between the investment-financing dataset and the comparison dataset when using Random Forest. However, a difference emerged with LightGBM. For LightGBM, incorporating network indicators resulted in worse performance.

Table 8. LightGBM true positive count by dataset configuration.

Actual	Predicted			
	Total	Financial	Investment-financing	Comparison
951	420	264	58	97

Resampling: Looking at the resampling results across six industries covering 317 bankrupt companies in total, SMOTE-ENN achieved the highest accuracy with 149 TPs, followed by SMOTE with 141 TPs, and K-means with 129 TPs (Table 9).

Table 9. LightGBM true positive count by resampling method and dataset configuration.

Resampling	Actual	Predicted			
		Total	Financial	Investment-financing	Comparison
SMOTE	317	141	85	21	35
SMOTE-ENN	317	149	98	22	29
K-means	317	129	81	15	33
Total	951	420	264	58	97

Industry-Specific Results: Table 10 presents the prediction results categorized by industry and dataset configuration. The financial dataset for the real estate industry produced the highest TP count with 62 companies, whereas the investment-financing dataset for the electrical equipment industry showed the lowest TP count of 0.

Table 10. LightGBM true positive count and recall rate by industry and dataset configuration.

Industry	Total			Financial			Investment-financing			Comparison		
	Actual	Predicted	Recall	Actual	Predicted	Recall	Actual	Predicted	Recall	Actual	Predicted	Recall

 doi:10.20944/preprints202510.2374.v1

Construction	105	61	58.10	75	47	62.70	15	6	40.00	15	8	53.17
Real estate	228	139	60.96	102	62	60.78	63	36	57.14	63	41	65.69
Service	78	34	43.95	48	24	49.28	15	2	10.00	15	9	60.87
Retail	129	59	45.67	69	47	68.00	30	8	26.67	30	4	13.33
Electrical equipment	276	67	24.30	126	50	39.58	75	0	0.00	75	17	22.92
Wholesale	135	59	43.86	69	35	50.72	33	7	21.21	33	17	52.17
Total	951	420	44.15	489	264	54.08	231	58	25.32	231	97	41.95

Resampling Results by Industry: Table 11 presents the resampling results for each industry. The highest TP count (49) was generated by the construction industry using SMOTE-ENN. Conversely, the lowest TP count (11) is observed for the service industry using both SMOTE-ENN and k-means.

Table 11. LightGBM true positive count and recall rate by resampling method and industry.

Industry	Total			SMOTE			SMOTE-ENN			K-means		
	Actual	Predicted	Recall	Actual	Predicted	Recall	Actual	Predicted	Recall	Actual	Predicted	Recall
Construction	105	61	58.10	35	19	54.29	35	21	60.00	35	21	60.00
Real estate	228	139	60.96	76	47	61.84	76	49	64.47	76	44	57.89
Service	78	34	43.95	26	12	46.15	26	11	42.31	26	11	42.31
Retail	129	59	45.67	43	21	48.83	43	22	51.16	43	17	39.53
Electrical equipment	276	67	24.30	92	24	26.09	92	28	30.43	92	15	16.30
Wholesale	135	59	43.86	45	19	42.22	45	19	42.22	45	21	46.67
Total	951	420	44.15	317	141	44.48	317	149	47.00	317	129	40.69

The combined results in Tables 10 and 11 show that, when using LightGBM, 139 TPs were generated for the real estate industry. By contrast, the lowest TP count (34) was observed in the service industry. Recall was highest in the real estate industry (60.96 %) and lowest in the wholesale industry (24.30 %).

Feature Importance: To identify the features that were significant for the model's predictions, we visualized the feature importance scores assigned by LightGBM. Figure 2 illustrates the top seven features (out of 173) for the electrical-equipment industry using the financial dataset processed using SMOTE-ENN.

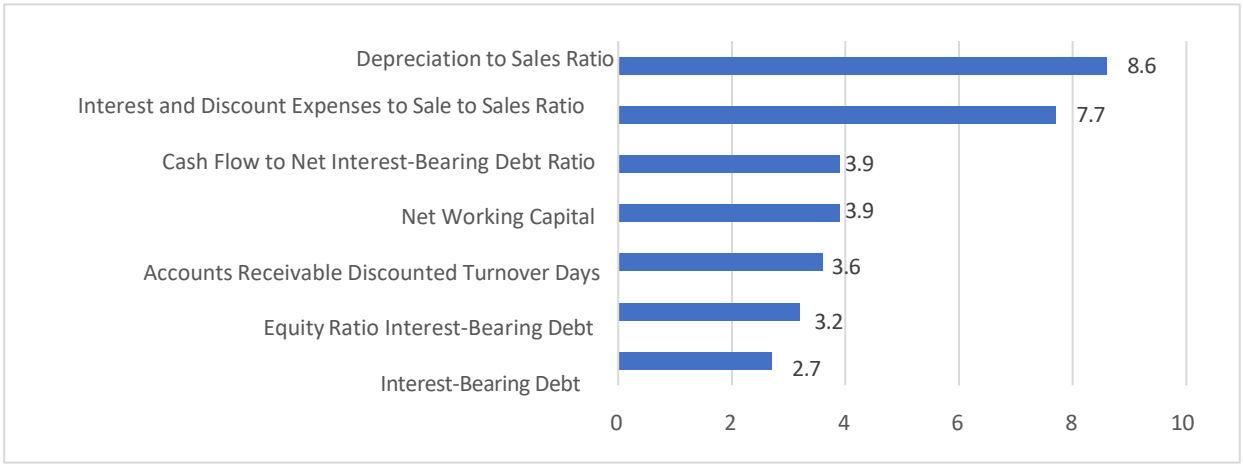


Figure 2. LightGBM feature importance scores for 173 features under SMOTE-ENN resampling using the financial dataset for the electrical equipment industry.

 doi:10.20944/preprints202510.2374.v1

3.2. Second-Stage Analysis Results (7 Features)

This study proposes a two-stage bankruptcy prediction model using financial data from listed Japanese companies. The first stage involved model training, using all 173 available financial indicators. The second stage improves prediction accuracy through passive dimensionality reduction to eliminate noise and prevent overfitting. The second stage applies feature selection using Random Forest and LightGBM to iteratively reduce the feature set from 173 to a smaller optimal set. The aim is to enhance the model's predictive performance by eliminating noise and irrelevant features, particularly when identifying bankrupt companies. The feature selection process followed a wrapper-based approach involving iterative model training and evaluation based on TPs and recall to identify the optimal number of features for each industry category. The procedure is outlined as follows:

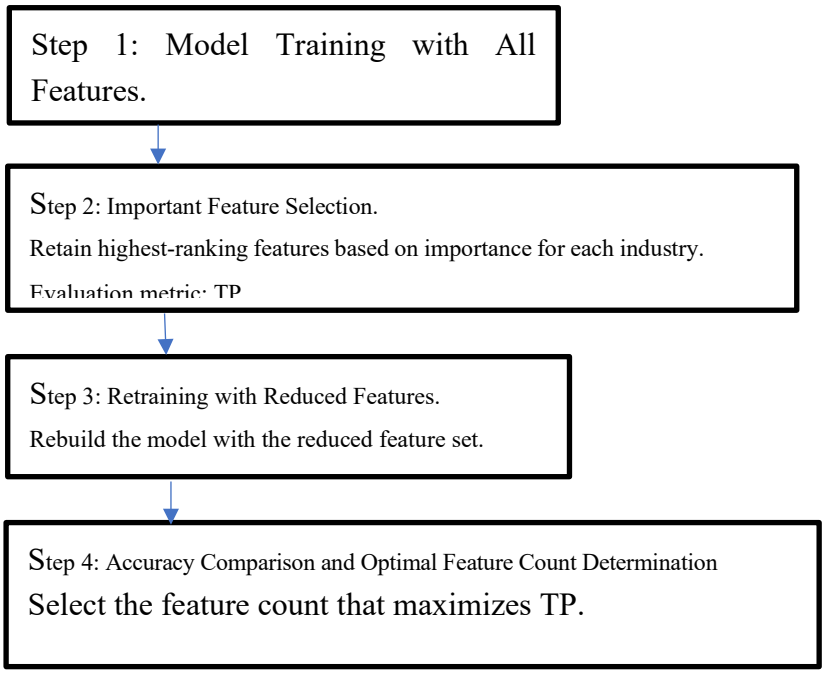
- Step 1: Train the model using all 173 features.
- Step 2: Based on performance (TP count), remove less important features and retain the top features for each industry.
- Step 3: Train a new model using the reduced feature set.
- Step 4: Steps 2 and 3 are repeated until the TP reaches its maximum. Point at which the TP peaks determine the optimal feature set. The results were as follows:
- Among the seven selected features, Random Forest achieved 566 TPs, whereas LightGBM achieved 451. Thus, the Random Forest model predicted 115 more bankruptcies.

In contrast, as previously mentioned, when all 173 features were used, Random Forest predicted 478 bankruptcies and LightGBM predicted 420. These results demonstrate that by employing the wrapper-based approach, which was the key objective of this study, we successfully reduced dimensionality, eliminated noise, and improved model performance.

Table 12 and Figure 3 show how the TP count changed as the number of features was progressively reduced from 173 to two through feature selection, representing the sum of the SMOTE, SMOTE+ENN, and k-means results.

Table 12. Changes in true positive count by feature count.

Features	4	5	7	15	130	161
TP	559	560	566	529	504	483



 doi:10.20944/preprints202510.2374.v1

Figure 3. Workflow of wrapper-based feature selection (Steps 1–4).

3.2.1. Random Forest

Table 13 presents the results generated using the 7-feature set. The total TP count across the financial, investment financing, and comparison datasets was 566. By contrast, as mentioned previously, when all 173 features were used, Random Forest achieved 478 TPs. This demonstrates that using the seven selected features predicted 88 more bankruptcies (566 vs. 478 TPs) with Random Forest than using all 173 features. Thus, we successfully achieved our study's objective of improving bankruptcy prediction accuracy. Considering the results by dataset configuration, the financial dataset, which had the largest sample size, showed the best prediction accuracy, with 303 TPs. The model produced 142 TPs with the investment financing dataset and 121 TPs with the comparison dataset, which demonstrates the added value of incorporating investment financing network indicators.

Table 13. Random Forest true positive counts by dataset configuration.

Actual	Predicted			
	Total	Financial	Investment-financing	Comparison
951	566	303	142	121

Resampling: SMOTE-ENN resampling achieved the highest accuracy, with 211 TPs across all three datasets combined (Table 14). K-means followed with 206 TPs and SMOTE with 149 TPs. In comparison, when all 173 features were used, SMOTE, SMOTE-ENN, and K-means achieved 115, 177, and 186 TPs, respectively, indicating improvements of 34, 34, and 20 additional TPs, respectively, with the seven-feature set compared with all 173 features.

Table 14. Random Forest true positive counts by resampling method and dataset configuration.

Resampling	Actual	Predicted			
		Total	Financial	Investment-financing	Comparison
SMOTE	317	149	84	34	31
SMOTE-ENN	317	211	113	52	46
K-means	317	206	106	56	44
Total	951	566	303	142	121

As shown in Table 15, the Random Forest model achieved the highest number of TPs when applied to the financial dataset of the construction industry. Specifically, it correctly identified 87 bankrupt firms out of 126 actual bankruptcies. In contrast, the comparative dataset exhibited the lowest performance in the service industry, with only two TPs.

Table 15. Random Forest true positive count and recall rate by industry and dataset configuration.

Industry	Total			Financial			Investment-financing			Comparison		
	Actual	Predicted	Recall	Actual	Predicted	Recall	Actual	Predicted	Recall	Actual	Predicted	Recall
Construction	276	172	62.32	126	87	69.05	75	44	58.67	75	41	54.67
Real estate	228	157	68.86	102	71	69.61	63	43	68.25	63	43	68.25
Service	105	57	54.29	75	52	69.33	15	3	20.00	15	2	13.33
Retail	129	71	55.04	69	45	65.22	30	14	46.67	30	12	40.00
Electrical equipment	276	61	62.32	126	18	69.05	75	29	58.67	75	14	54.67
Wholesale	78	48	61.54	48	30	62.50	50	9	60.00	15	9	60.00

 doi:10.20944/preprints202510.2374.v1

Total	951	566	59.52	489	303	61.96	231	142	61.47	231	121	52.38

Resampling Results by Industry: Table 16 presents the resampling results for each industry. In the construction industry, SMOTE-ENN yields the highest TP count of 67. Conversely, for the wholesale industry, using SMOTE resulted in the lowest TP count at only 12.

The combined results in Tables 15 and 16 indicate that the highest TP count (172) was recorded for the construction industry, whereas the lowest TP count (48) was recorded for the wholesale industry. In terms of recall, the highest rate (68.86%) was observed for the real estate industry data, whereas the lowest rate (54.29%) was observed for the service industry data.

Table 16. Random Forest true positive count and recall rate by resampling method and industry.

Industry	Total			SMOTE			SMOTE-ENN			K-means		
	Actual	Predicted	Recall	Actual	Predicted	Recall	Actual	Predicted	Recall	Actual	Predicted	Recall
Construction	276	172	62.32	92	43	46.74	92	67	72.83	92	62	67.39
Real estate	228	157	68.86	76	46	60.53	76	60	78.95	76	51	67.11
Service	105	57	54.29	35	16	45.71	35	21	60.00	35	20	57.14
Retail	129	71	55.04	43	18	41.86	43	26	60.47	43	27	62.79
Electrical equipment	135	61	45.19	45	14	31.11	45	20	44.44	45	27	60.00
Wholesale	78	48	61.54	26	12	46.15	26	17	65.38	26	19	73.08
Total	951	566	59.52	317	149	47.00	317	211	66.56	317	206	64.98

Feature Importance: Table 17 compares the top seven features identified by Random Forest from the full 173-feature set with the seven features selected for the optimized model. This table highlights the top seven features selected from the original 173 features for comparison.

Notably, no network indicators were selected among the final seven features, and all the selected features were financial indicators. Furthermore, Cash Flow to Net Interest-Bearing Debt Ratio and Net Interest Burden to Sales Ratio appear among the seven selected features and among the top seven indicators of the 173-feature set, demonstrating their importance for bankruptcy prediction with Random Forest. Moreover, when using the seven-feature set, two cash flow indicators are included: cash flow to net interest-bearing debt ratio and operating cash flow to sales ratio. When using all 173 features, four of the top seven features are cash flow-related: the Cash Flow to Net Interest-Bearing Debt Ratio, Dividends to Cash Flow Ratio, Cash Flow to Long-Term Debt Ratio, and Cash Flow to Fixed Liabilities Ratio. Given that cash shortages are often the primary cause of corporate bankruptcy, this finding indicates that the model successfully identifies the relevant features.

Table 17. Comparison of the top seven features identified by Random Forest from the full set with the seven features selected for the optimized model.

173 Features	7 Features
Cash Flow to Net Interest-Bearing Debt Ratio	Net Profit Margin
Net Working Capital	Cash Flow to Net Interest-Bearing Debt Ratio
Dividends to Cash Flow Ratio	Equity Ratio
Cash Flow to Long-Term Debt Ratio	Net Interest Burden to Sales Ratio
Net Interest Burden to Sales Ratio	Equity Growth Rate (Year-on-Year)
Interest and Discount Expenses to Sales Ratio	Operating Cash Flow to Sales Ratio
Cash Flow to Fixed Liabilities Ratio	Payment Reserve Ratio

For the electrical equipment industry, Figure 4 presents the importance scores of the top seven features computed on the financial dataset following SMOTE-ENN resampling.

 doi:10.20944/preprints202510.2374.v1

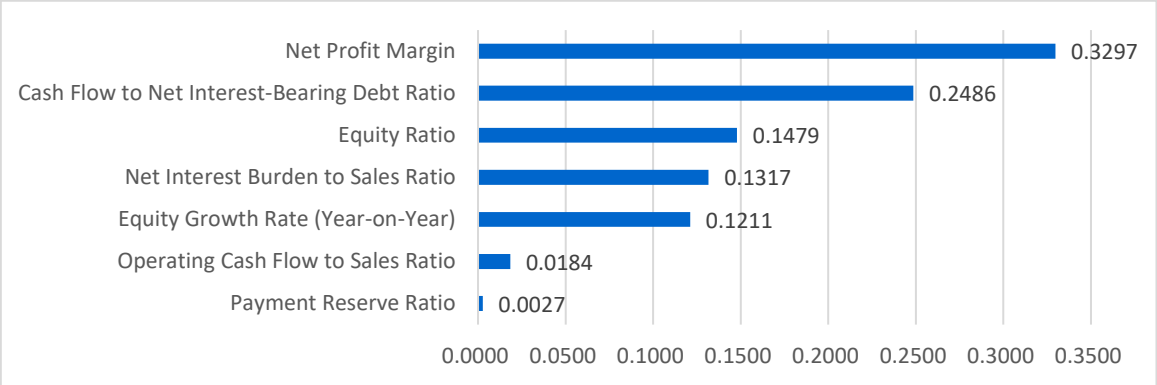


Figure 4. Importance scores for the top seven features under SMOTE-ENN resampling using the financial dataset for the electrical equipment industry.

3.2.2. LightGBM

Table 18 presents the results obtained using these seven features. Among the datasets, the highest prediction accuracy for 283 TPs was achieved using the financial dataset with the largest sample size.

Table 18. LightGBM true positive count by dataset configuration.

Actual	Predicted			
	Total	Financial	Investment-financing	Comparison
951	451	283	87	81

Resampling: Across all six industries and three dataset configurations (encompassing 317 bankrupt companies in total), SMOTE resampling achieved the highest accuracy with 199 TPs, followed by SMOTE-ENN with 191 TPs, and k-means with 61 TPs.

Table 19. LightGBM true positive count by resampling method and dataset configuration.

Resampling	Actual	Predicted			
		Total	Financial	Investment-financing	Comparison
SMOTE	317	199	116	44	39
SMOTE-ENN	317	191	114	41	36
K-means	317	61	53	2	6
Total	951	451	281	87	81

Industry-Specific Results :As shown in Table 20, for LightGBM, the highest TP count (72) is achieved using the financial dataset for the real estate industry, whereas the lowest TP count of 2 is observed when using the investment-financing dataset for the retail industry.

Table 20. LightGBM true positive count by industry and dataset configuration.

Industry	Total			Financial			Investment-financing			Comparison		
	Actual	Predicted	Recall	Actual	Predicted	Recall	Actual	Predicted	Recall	Actual	Predicted	Recall
Construction	105	72	68.57	75	53	72.60	15	10	66.67	15	9	60.00
Real estate	228	138	60.53	102	72	70.59	63	34	53.97	63	32	50.79
Service	78	42	53.85	48	32	66.67	15	5	33.33	15	5	33.33
Retail	129	44	34.11	69	34	49.28	30	2	6.67	30	8	26.67
Electrical	276	93	33.70	126	53	42.06	75	25	33.33	75	15	20.00

 doi:10.20944/preprints202510.2374.v1

equipment												
Wholesale	135	62	45.93	69	37	53.62	33	12	36.36	33	13	39.39
Total	951	451	47.42	489	281	57.46	231	88	38.10	231	82	35.50

Resampling Results by Industry: As shown in Table 21, which presents the LightGBM resampling results by industry, SMOTE-ENN yielded 60 TPs with data for the real estate industry. K-means showed the worst performance with the electrical equipment industry data at zero TPs.

Table 21. LightGBM true positive count and recall rate by resampling method and industry.

Industry	Total			SMOTE			SMOTE -ENN			K-means		
	Actual	Predicted	Recall	Actual	Predicted	Recall	Actual	Predicted	Recall	Actual	Predicted	Recall
Construction	105	72	68.57	35	28	80.00	35	28	80.00	35	17	48.57
Real estate	228	138	60.53	76	59	77.63	76	60	78.95	76	19	25.00
Service	78	43	55.13	26	15	57.69	26	16	61.65	26	2	7.69
Retail	129	43	33.33	43	19	44.19	43	22	51.16	43	3	6.98
Electrical equipment	276	93	33.70	92	54	58.70	92	39	42.39	92	0	0.00
Wholesale	135	62	45.93	45	25	55.56	45	27	60.00	45	10	22.22
Total	951	451	47.42	317	199	62.78	317	191	60.25	317	61	19.24

As the combined results in Tables 20 and 21 show, when the seven-feature LightGBM model is applied across industries, the highest TP count (138) is predicted for the real estate industry. Conversely, the lowest TP count of 43 is predicted for the service industry. In terms of recall, the highest rate was recorded for the construction industry (68.57 %), whereas the lowest rate was observed for the wholesale industry (33.70 %).

Feature Importance: Table 22 presents a comparison of the top seven features identified by LightGBM from the full 173-feature set with the seven features selected for the optimized model. The Depreciation to Sales ratio feature appears in both the seven selected features and the top seven of the full set, underscoring its critical role in LightGBM-based bankruptcy pre- diction. In the seven-feature model, four cash flow metrics were selected (dividend-to-free cash flow ratio, instant coverage cash flow, cash flow to debt ratio, and cash flow to current liabilities ratio), whereas in the full 173-feature ranking, the Cash Flow to Net Interest-Bearing Debt Ratio also emerged among the top indicators. Given that insufficient cash flow is the leading cause of corporate bankruptcies, these results confirm that the proposed approach effectively identified relevant features.

Table 22. Comparison of the top seven features identified by LightGBM from the full set with the seven features selected for the optimized model.

173 Features	7 Features
Depreciation to Sales Ratio	Cash and Deposits to Interest-Bearing Debt Ratio
Interest and Discount Expenses to Sales Ratio	Depreciation to Sales Ratio
Net Working Capital	Dividends to Free Cash Flow Ratio
Cash Flow to Net Interest-Bearing Debt Ratio	SG&A to Sales Ratio
Accounts Receivable Discounted Turnover Days	Instant Coverage Cash Flow
Equity Ratio	Cash Flow to Debt Ratio
Interest-Bearing Debt	Cash Flow to Current Liabilities Ratio

Figure 5 illustrates the top seven features (out of 173) for the electrical equipment industry when using a financial dataset processed with SMOTE-ENN.

 doi:10.20944/preprints202510.2374.v1

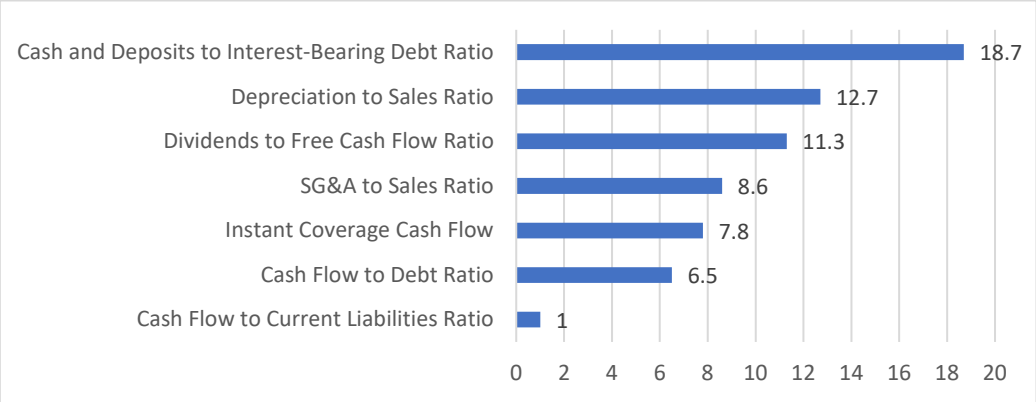


Figure 5. Importance scores for the top seven features under SMOTE-ENN resampling using the financial dataset for the electrical equipment industry.

4. Discussion

This study contributes significantly to the field of bankruptcy prediction by proposing a novel two-stage machine learning approach that integrates ensemble learning, feature selection, and industry-specific analysis. By employing Random Forest and LightGBM, this research addresses critical challenges in bankruptcy prediction, including handling high-dimensional financial data, imbalanced datasets, and the need for industry-specific modeling. The innovative wrapper-based feature selection process effectively reduced the dimensionality of the dataset from 173 financial indicators to seven optimal features, thereby eliminating noise and enhancing model performance.

Furthermore, we construct industry-specific models for six sectors based on the Tokyo Stock Exchange classification, uncovering unique bankruptcy patterns and causes within each industry. This approach not only improves the generalizability of the models but also provides actionable insights for stakeholders in different industries. The research also demonstrated the superior performance of Random Forest over LightGBM, with the former achieving 566 true positives using a seven-feature set, an improvement of 88 cases compared with the full feature set.

By addressing the challenges of imbalanced data through advanced resampling techniques such as SMOTE, SMOTE-ENN, and k-means, this study ensures a robust model the performance across diverse datasets. These methodological innovations have contributed to academic research and practical applications by offering a comprehensive framework for accurate and interpretable bankruptcy prediction. These findings underscore the importance of tailored feature selection and industry-specific analysis, paving the way for future advancements in predictive modeling and decision support systems.

Author Contributions: Conceptualization, M.M. and H.K.; methodology, M.M.; software, M.M.; validation, M.M. and H.K.; formal analysis, M.M.; investigation, M.M.; resources, M.M.; data curation, M.M.; writing—original draft preparation, M.M.; writing—review and editing, M.M. and H.K.; visualization, M.M.; supervision, M.M.; project administration, M.M.; funding acquisition, H.K.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

 doi:10.20944/preprints202510.2374.v1

Abbreviations

The following abbreviations are used in this manuscript:

RF	Random Forest
LGBM	Light Gradient Boosting Machine
TP	True Positive
FP	False Positive
TP	True Positive
FN	False Negative
TN	True Negative
SMOTE	Synthetic Minority Oversampling Technique
ENN	Edited Nearest Neighbors
KMeans	K-Means Clustering
TSE	Tokyo Stock Exchange

References

1. Alaka, H.A., Oyedele, L.O., Owolabi, H.A., Kumar, V., Ajayi, S.O., Akinade, O.O., and Bilal, M. (2018). Systematic review of bankruptcy prediction models: Towards a framework for tool selection. *Expert Systems with Applications*, 94, 164–184. https://doi.org/10.1016/j.eswa.2017.10.040
2. Altman, E.I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23, 589–609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x
3. Altman, E.I., Iwanicz-Drozdowska, M., Laitinen, E.K., and Suvas, A. (2017). Financial distress prediction in an international context: A review and empirical analysis of Altman's Z-score model. *Journal of International Financial Management and Accounting*, 28, 131–171. https://doi.org/10.1111/jifm.12053
4. Ansah-Narh, T., Nortey, E.N.N., Proven-Adzri, E., and Opoku-Sarkodie, R. (2024). Enhancing corporate bankruptcy prediction via a hybrid genetic algorithm and domain adaptation learning architecture. *Expert Systems with Applications*, 258, 120654. https://doi.org/10.1016/j.eswa.2024.125133
5. Atiya, A.F. (2001). Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on Neural Networks*, 12, 929–935. https://doi.org/10.1109/72.935101
6. Bandyopadhyay, S., and Lang, M. (2013). Ensemble learning for financial default prediction. *Journal of Finance and Data Science*, 1, 69–81.
7. Batista, G.E.A.P.A., Prati, R.C., and Monard, M.C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6, 20–29. https://doi.org/10.1145/1007730.1007735
8. Beaver, W.H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, 4, 71–111. https://doi.org/10.2307/2490171
9. Bellovary, J.L., Giacomino, D.E., and Akers, M.D. (2007). A review of bankruptcy prediction studies: 1930 to present. *Journal of Financial Education*, 33, 1–42.
10. Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. https://doi.org/10.1023/A:1010933404324
11. Buda, M., Maki, A., and Mazurowski, M.A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011
12. Chandra, D.K., Ravi, V., and Bose, I. (2009). Failure prediction of dotcom companies using hybrid intelligent techniques. *Expert Systems with Applications*, 36, 4830–4837. https://doi.org/10.1016/j.eswa.2008.05.047
13. Chava, S., and Jarrow, R.A. (2004). Bankruptcy prediction with industry effects. *Review of Finance*, 8, 537–569. https://doi.org/10.1093/rof/8.4.537
14. Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. https://doi.org/10.1613/jair.953
15. Chen, T., and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

 doi:10.20944/preprints202510.2374.v1

16. Choi, I., and Lee, J. (2018). Multi-label learning for corporate bankruptcy prediction. *Decision Support Systems*, 110, 87–98.
17. Ciampi, F. (2015). Corporate governance characteristics and default prediction modeling for small enterprises: An empirical analysis of Italian firms. *Journal of Business Research*, 68, 1012–1025. https://doi.org/10.1016/j.jbusres.2014.10.003
18. Dasilas, A., and Rigani, A. (2024). Machine learning techniques in bankruptcy prediction: A systematic literature review. *Expert Systems with Applications*, 255, 124761. https://doi.org/10.1016/j.eswa.2024.124761
19. Dietterich, T.G. (2000). Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems* (pp. 1–15).
20. Dikshit, A., and Pradhan, B. (2021). Interpretable and explainable AI (XAI) model for spatial drought prediction. *Science of the Total Environment*, 801, 149797. https://doi.org/10.1016/j.scitotenv.2021.149797
21. Du Jardin, P. (2016). A two-stage classification technique for bankruptcy prediction. *European Journal of Operational Research*, 254, 236–252. https://doi.org/10.1016/j.ejor.2016.03.008
22. Fedorova, E., Ledyaeva, S., Drogovoz, P., and Nevredinov, A. (2022). Economic policy uncertainty and bankruptcy filings. *International Review of Financial Analysis*, 82, 102174. https://doi.org/10.1016/j.irfa.2022.102174
23. Fernandez, A., García, S., and Herrera, F. (2018). A survey on imbalanced classification in credit scoring: SMOTE-based techniques. *Pattern Recognition*, 91, 346–362.
24. Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., and Herrera, F. (2018). *Learning from imbalanced datasets*. Springer. https://doi.org/10.1007/978-3-319-98074-4
25. Ganaie, M.A., and Hu, M. (2022). Ensemble deep learning: A review. *Knowledge-Based Systems*, 239, 108098.
26. García, V., Marqués, A.I., Sánchez, J.S., and Ochoa-Domínguez, H.J. (2019). Dissimilarity-based linear models for corporate bankruptcy prediction. *Computational Economics*, 53, 1019–1031. https://doi.org/10.1007/s10614-017-9783-4
27. Giudici, P., and Hadji-Misheva, B. (2022). Explainable ML for credit scoring and bankruptcy prediction. *Risks*, 10, 104.
28. Guillén, M., and Salas, A. (2021). Bankruptcy prediction combining feature selection and machine learning classifiers. *Sustainability*, 13, 6436.
29. Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157–1182.
30. He, H., and Garcia, E.A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21, 1263–1284. https://doi.org/10.1109/TKDE.2008.239
31. Heo, J., and Yang, J.Y. (2014). AdaBoost-based bankruptcy forecasting of Korean construction companies. *Applied Soft Computing*, 24, 494–499. https://doi.org/10.1016/j.asoc.2014.08.009
32. Hernandez Tinoco, M.H., and Wilson, N. (2013). Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables. *International Review of Financial Analysis*, 30, 394–419. https://doi.org/10.1016/j.irfa.2013.02.013
33. Hillegeist, S.A., Keating, E.K., Cram, D.P., and Lundstedt, K.G. (2004). Assessing the probability of bankruptcy. *Review of Accounting Studies*, 9, 5–34. https://doi.org/10.1023/B:RAST.0000013627.90884.b7
34. Hossari, G., and Rahman, S. (2022). Artificial intelligence and bankruptcy prediction: The relevance of feature selection. *International Journal of Finance and Economics*, 27, 2103–2122.
35. Huang, J., and Ling, C.X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17, 299–310. https://doi.org/10.1109/TKDE.2005.50
36. Huang, Z., Chen, H., Hsu, C.-J., Chen, W.-H., and Wu, S. (2004). Credit rating analysis using support vector machines and neural networks: A comparative market study. *Decision Support Systems*, 37, 543–558. https://doi.org/10.1016/S0167-9236(03)00086-1
37. Iparraguirre-Villanueva, O., and Cabanillas-Carbonell, M. (2023). Predicting business bankruptcy: A comparative analysis with machine learning models. *Economies*, 11, 122.
38. Jabeur, S.B., Gharib, C., Mefteh-Wali, S., and Ben Arfi, W. (2021). CatBoost model and artificial intelligence techniques for corporate failure prediction. *Technological Forecasting and Social Change*, 166, 120658. https://doi.org/10.1016/j.techfore.2021.120658

 doi:10.20944/preprints202510.2374.v1

39. Jain, V., and Johnson, R. (2020). Feature selection and ensemble learning for bankruptcy prediction. *Journal of Risk and Financial Management*, 13, 210.
40. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 3146–3154.
41. Kim, A., and Yoon, S. (2023). Corporate bankruptcy prediction with domain-adapted BERT. *arXiv preprint*, arXiv:2312.03194.
42. Kim, H.Y. (2011). Bankruptcy prediction using support vector machine with optimal choice of kernel function and regularization parameters. *Expert Systems with Applications*, 38, 511–517.
43. Kim, H., Cho, H., and Ryu, D. (2022). Corporate bankruptcy prediction using machine learning methodologies with a focus on sequential data. *Computational Economics*, 59, 1231–1249. https://doi.org/10.1007/s10614-021-10126-5
44. Ko, B., Kim, D., and Kang, B. (2021). A hybrid feature selection approach for bankruptcy prediction. *Applied Intelligence*, 51, 111–128.
45. Kotze, M., and Beukes, C. (2021). Feature extraction using hybrid genetic algorithms and XGBoost. *Neurocomputing*, 452, 111–122.
46. Kraus, C., and Feuerriegel, S. (2019). Decision support for bankruptcy prediction using machine learning: A comparison of boosting and bagging. *Decision Support Systems*, 120, 113–126.
47. Lahsasna, A., Ainon, R.N., and Wah, T.Y. (2018). Business failure prediction using ensemble machine learning. *International Journal of Advanced Computer Science and Applications*, 9, 45–52.
48. Lee, C.-C., and Chen, M.-L. (2020). Ensemble models for predicting corporate financial distress: A performance comparison. *Expert Systems with Applications*, 139, 112–124.
49. Lee, S., and Choi, W.S. (2013). A multi-industry bankruptcy prediction model using back-propagation neural network and multivariate discriminant analysis. *Expert Systems with Applications*, 40, 2941–2946. https://doi.org/10.1016/j.eswa.2012.12.009
50. Li, H., Sun, J., and Wu, J. (2022). Financial distress prediction using attention-based deep learning. *Expert Systems with Applications*, 199, 116–137.
51. Lin, C.T., Chiu, C.C., and Tsai, C.Y. (2010). A hybrid neural network model for credit scoring. *International Journal of Electronic Business Management*, 8, 254–261.
52. Lin, F.Y., and McClean, S. (2001). A data mining approach to the prediction of corporate failure. *Knowledge-Based Systems*, 14, 189–195. https://doi.org/10.1016/S0950-7051(01)00096-X
53. Liu, X.-Y., Wu, J., and Zhou, Z.-H. (2009). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics – Part B*, 39, 539–550. https://doi.org/10.1109/TSMCB.2008.2007853
54. Liu, Y., and Wu, H. (2021). Bankruptcy prediction using SMOTE-ENN and LightGBM. *Sustainability*, 13, 8021.
55. Lundberg, S.M., and Lee, S.I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 4765–4774.
56. Martín, R., and Rojo, A. (2019). Bankruptcy prediction using time series of financial ratios. *Applied Economics Letters*, 26, 1605–1609.
57. Mishra, D., and Singh, A. (2022). Predicting bankruptcy using XGBoost and SMOTE. *Journal of Risk and Financial Management*, 15, 142.
58. Mselmi, N., Lahiani, A., and Hamza, T. (2017). Financial distress prediction: The case of French SMEs. *International Review of Financial Analysis*, 50, 67–80. https://doi.org/10.1016/j.irfa.2017.02.004
59. Nam, J., and Jinn, T. (2000). Bankruptcy prediction: Evidence from Korean listed companies during the IMF financial crisis. *Journal of International Financial Management and Accounting*, 11, 178–197. https://doi.org/10.1111/1467-646X.00061
60. Narvekar, A., and Guha, D. (2021). Bankruptcy prediction using machine learning and an application to the case of the COVID-19 recession. In *Data Science in Finance and Economics*, 1, 180–195. https://doi.org/10.3934/DSFE.2021010

 doi:10.20944/preprints202510.2374.v1

61. Ohlson, J.A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18, 109–131. https://doi.org/10.2307/2490395
62. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
63. Pramodh, T.R., and Ravi, V. (2016). Rule extraction and feature selection techniques – A review. *International Journal of Computers and Applications*, 975, 8887.
64. Radovanovic, J., and Haas, C. (2023). The evaluation of bankruptcy prediction models based on socio-economic costs. *Expert Systems with Applications*, 227, 120275. https://doi.org/10.1016/j.eswa.2023.120275
65. Rashid, A., Hassan, M.K., and Karamat, H. (2021). Firm size and the interlinkages between sales volatility, exports, and financial stability of Pakistani manufacturing firms. *Eurasian Business Review*, 11, 111–134. https://doi.org/10.1007/s40821-020-00162-w
66. Ravi Kumar, P., and Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review. *European Journal of Operational Research*, 180, 1–28. https://doi.org/10.1016/j.ejor.2006.08.043
67. Ravisankar, P., and Ravi, V. (2010). Financial distress prediction in banks using GMDH, CP-NN, and fuzzy ARTMAP. *Knowledge-Based Systems*, 23, 823–831. https://doi.org/10.1016/j.knosys.2010.05.007
68. Razzak, I., Imran, M., and Xu, G. (2019). Deep learning for credit scoring: A review. *Information Processing and Management*, 56, 102–128.
69. Sáez, J.A., Luengo, J., Stefanowski, J., and Herrera, F. (2015). SMOTE-IPF: A filtering method to pre-process data. *Information Sciences*, 291, 184–203. https://doi.org/10.1016/j.ins.2014.08.051
70. Sánchez-Medina, A.J., Blázquez-Santana, F., Cerviño-Cortínez, D.L., and Pellejero, M. (2024). Ensemble methods for bankruptcy resolution prediction: A new approach. *Computational Economics*.
71. Shetty, S., Musa, M., and Brédart, X. (2022). Bankruptcy prediction using machine learning techniques. *Journal of Risk and Financial Management*, 15, 35. https://doi.org/10.3390/jrfm15010035
72. Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The Journal of Business*, 74, 101–124. https://doi.org/10.1086/209665
73. Succurro, M., Arcuri, G., and Costanzo, G.D. (2019). A combined approach based on robust PCA to improve bankruptcy forecasting. *Review of Accounting and Finance*, 18, 296–320. https://doi.org/10.1108/RAF-04-2018-0077
74. Sun, J., Li, H., Huang, Q.-H., and He, K.-Y. (2014). Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. *Knowledge-Based Systems*, 57, 41–56. https://doi.org/10.1016/j.knosys.2013.12.006
75. Sun, L., and Shenoy, P.P. (2007). Using Bayesian networks for bankruptcy prediction: Some methodological issues. *European Journal of Operational Research*, 180, 738–753. https://doi.org/10.1016/j.ejor.2006.04.019
76. Tang, L., and Yan, H. (2021). Financial distress prediction based on stacking ensemble. *Journal of Intelligent and Fuzzy Systems*, 41, 3147–3159.
77. Tang, Y., Zhang, Y.-Q., and Chawla, N.V. (2009). SVMs modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man and Cybernetics*, 39, 281–288.
78. Tsai, C.-F., and Hsu, Y.-F. (2013). A meta-learning framework for credit scoring. *Expert Systems with Applications*, 40, 5124–5130.
79. Tsai, C.-F. (2009). Feature selection in bankruptcy prediction. *Knowledge-Based Systems*, 22, 120–127. https://doi.org/10.1016/j.knosys.2008.08.002
80. Varetto, F. (1998). Genetic algorithms applications in the analysis of insolvency risk. *Journal of Banking and Finance*, 22, 1421–1439. https://doi.org/10.1016/S0378-4266(98)00059-4
81. Wang, G., Ma, J., Huang, L., and Xu, K. (2010). Two credit scoring models based on dual strategy ensemble trees. *Knowledge-Based Systems*, 23, 899–908.
82. Wu, J., and Wang, Y. (2022). Corporate bankruptcy prediction using explainable boosting machine. *Expert Systems with Applications*, 187, 115859.
83. Wu, Y., Gaunt, C., and Gray, S. (2010). A comparison of alternative bankruptcy prediction models. *Journal of Contemporary Accounting and Economics*, 6, 34–45. https://doi.org/10.1016/j.jcae.2010.04.002

 doi:10.20944/preprints202510.2374.v1

84. Xu, Y., and Ouenniche, J. (2018). Slacks-based DEA and cross-benchmarking framework to evaluate bankruptcy predictive models. *Expert Systems with Applications*, 104, 240–253.
85. Yeh, C.-H., Chi, D.-J., and Lin, Y.-R. (2022). Improved bankruptcy prediction using feature selection and classification algorithms. *Journal of Risk and Financial Management*, 15, 329.
86. Yeh, I.-C., and Lien, C.-H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36, 2473–2480. https://doi.org/10.1016/j.eswa.2007.12.020
87. Yu, L., and Zhang, Z. (2020). A novel stacking ensemble learning framework for bankruptcy prediction. *IEEE Access*, 8, 58828–58840.
88. Zhang, Z., and Wang, J. (2021). Explainable deep learning model for financial distress prediction. *Expert Systems with Applications*, 185, 115655.
89. Zhao, Y., and Wang, G. (2018). Predicting corporate bankruptcy using ensemble learning and data balancing techniques. *Applied Soft Computing*, 72, 362–375.
90. Zmijewski, M.E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research*, 22, 59–82. https://doi.org/10.2307/2490859