Corporate Bankruptcy Prediction Using Machine Learning

Jialu Chen*

Department of Quantitative Finance, The Chinese University of Hong Kong, Hong Kong, China

*Corresponding author: 1155210991@link.cuhk.edu.hk

Abstract. As the advancement of machine learning, hundreds of methods have been proposed in solving corporate bankruptcy prediction problems. Facing with these options, decision-makers must decide the most effective methods. On the other hand, bankruptcy data usually involve high-dimensionality and extreme class imbalance, which may undermine the performance of classical models. This research designs and validates a systematic framework to address these issues, combining feature selection, sample rebalancing and machine learning model selection together. The study chooses five typical machine learning models, including Logistic Regression (LR), Support Vector machine (SVM), Decision Tree (DT), XGBoost (XGB) and Random Forest (RF). Also, this study designs the training and test data sets using a two-layered feature selection method, comparing two resampling methods and five classic machine learning models to form a final improved voting classifier with hyperparameters tuned. The experimental results confirm that XGB and RF, when combined with oversampling method, can provide the most robust prediction.

Keywords: Bankruptcy; Machines Learning; Feature Selection; Resampling; Voting Classifier.

1. Introduction

Corporate bankruptcy prediction is a core component of firm risk management, where an inaccurate prediction might lead to severe consequences such as credit defaults and investment losses. Therefore, the development of a robust model to predict company financial health is of significance. According to the study of Rybarova et al., bankruptcy prediction model is a system providing insights for detecting company financial health situation based on the selected indicators [1].

However, researchers may face two persistent challenges in this domain. First, bankruptcy data usually involve high dimensional features (ROA, liability-to-equity ratio, etc.), which may lead to model overfitting and computational inefficiency. Second, bankruptcy data are extremely imbalanced, with bankrupt companies accounting for only a small proportion. This might undermine traditional model performance that are generally trained to optimize the accuracy.

To solve these problems, recent research has shifted from traditional statistical methods, such as Altman's Z-score [2], to more advanced machine learning models, which have already demonstrated better performance in classification problems [3]. However, the optimal strategy for data pre-processing and model selection is still under discussion. For instance, while some studies have showed that SVM performs better in handling data with non-linear feature relationships or high dimensionality [4], it is also doubted that the black-box nature of SVM makes interpretation difficult, and also its high computational cost makes it inefficient [5]. Besides SVM, ensemble methods, especially XGB, have shown great promise. A study studying banks in the Europe by Climent et al. found that XGB outperformed other machine learning models including RF and LR, in dealing with bankruptcy prediction [6]. Although these studies showed the promise of machine learning, the optimal strategy of feature selection and model choice is inconclusive. For example, Liang et al. investigated multiple methods but concluded that no single combination was universally superior, encouraging people to use hybrid methods for further investigation [7].

In recent years, the field has gained more insights into the following three major trends. First, Deep Learning (DL), like Long Short-Term Memory (LSTM), is being explored widely. LSTM can analyze time series data, which can detect problems earlier [8]. Second, Explainable AI (XAI) uses methods like SHapley Additive exPlanations (SHAP) to interpret machine learning models like XGB, which helps to explain the reason why a model makes the decision [9]. Third, more advanced methods like Generative Adversarial Networks (GANs) have been proposed to deal with the imbalanced data set,

outperforming traditional sampling methods like SMOTE [10]. While these models are promising, they might involve huge computational complexity and are usually hard to interpret. Therefore, building an efficient method using traditional machine learning models remains important, which can reduce the computational complexity and they are also the basic benchmark for complex DL models.

This study aims to build an efficient bankruptcy prediction framework by solving the problems of high dimensionality and imbalanced data. To handle the dimensionality, the paper proposes a two-layered feature selection method: The researcher firstly selects features based on their feature importance scores with a Random Forest Classifier, and then further filters out features which have high correlations using the Pearson Correlation Coefficient. To deal with the imbalanced data, the researcher conducts a comparative analysis of two widely used resampling methods: SMOTE for oversampling and RandomUnderSampler (RUS) for undersampling.

The study compares five classic machine learning models: LR, SVM, DT, RF and XGB. Then to enhance the prediction stability, the research builds a soft-voting classifier with the best two performed models above (RF and XGB). Given the higher cost of failing to predict a bankruptcy, the researcher prioritizes recall and the F2-Score as the main evaluation metrics over the traditional choice of accuracy. This comprehensive framework aims to provide a reliable and efficient way for financial institutions to track bankruptcy risks.

2. Method

Data set for this research is collected from Kaggle, which contains 6819 companies in Taiwan with 95 financial features during the period of 1999-2009. The data involve 6599 non-bankrupt companies with only 220 bankrupt ones, accounting for only 3.2%, which will result in high false negative rates. To achieve better performance, feature selection and resampling methods are performed to get a more balanced data set before model fitting.

2.1. Feature Selection

Feature selection involves two steps. First, the research utilizes a Random Forest Classifier to calculate the feature importance score for all the 95 features. The rationale is that features with higher importance scores have more contribution to the model on distinguishing between bankrupt and non-bankrupt companies. As shown in Figure 1, a threshold of feature importance > 0.015 is set to select features with higher importance, with only 14 features remained.

After initial selection, as shown in Figure 2, Pearson correlation analysis is used to the features remained in the last step to avoid redundancy. Pearson correlation coefficients range from -1 to 1, where absolute values closer to 1 indicate stronger linear relationships. For each pair of features whose correlation coefficient > 0.8, the feature with lower random forest feature importance score is removed. This ensures that the retained feature retains higher power of prediction while also avoiding redundancy. For example, in our data set "Net profit before tax/paid-in capital" and "Persistent EPS in the last four seasons" has a correlation of 0.96 and the former one with importance score lower than the latter is excluded.

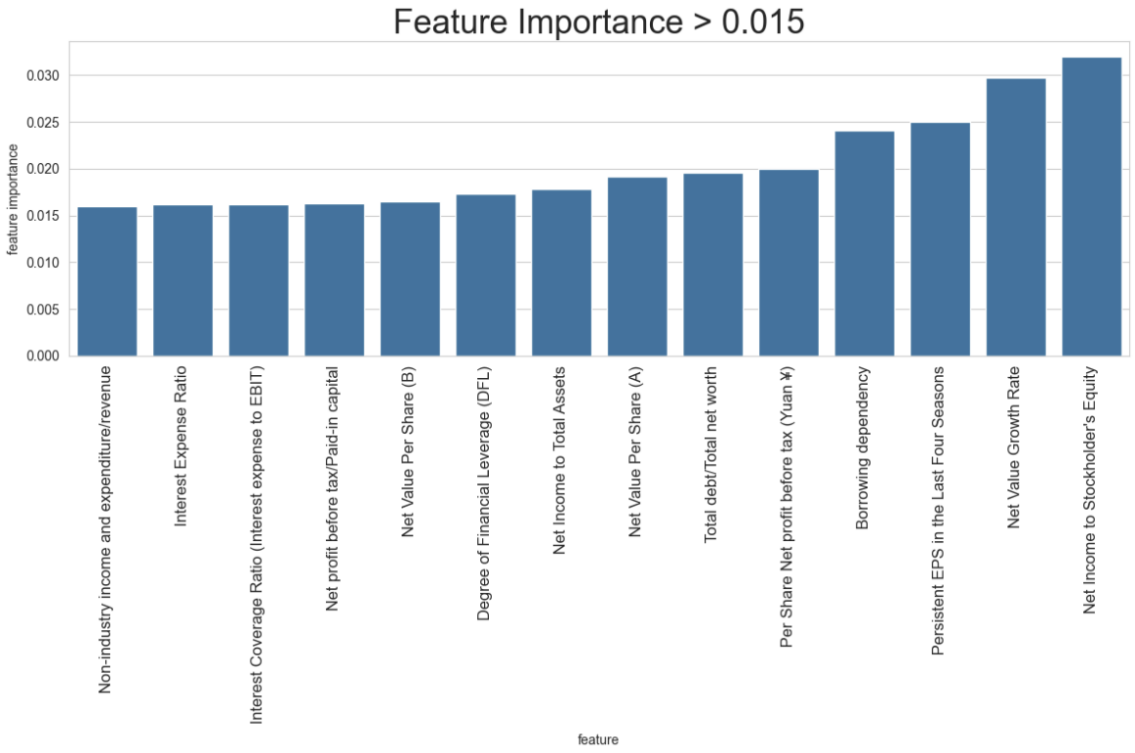

Fig. 1 Feature with Importance Score > 0.0015

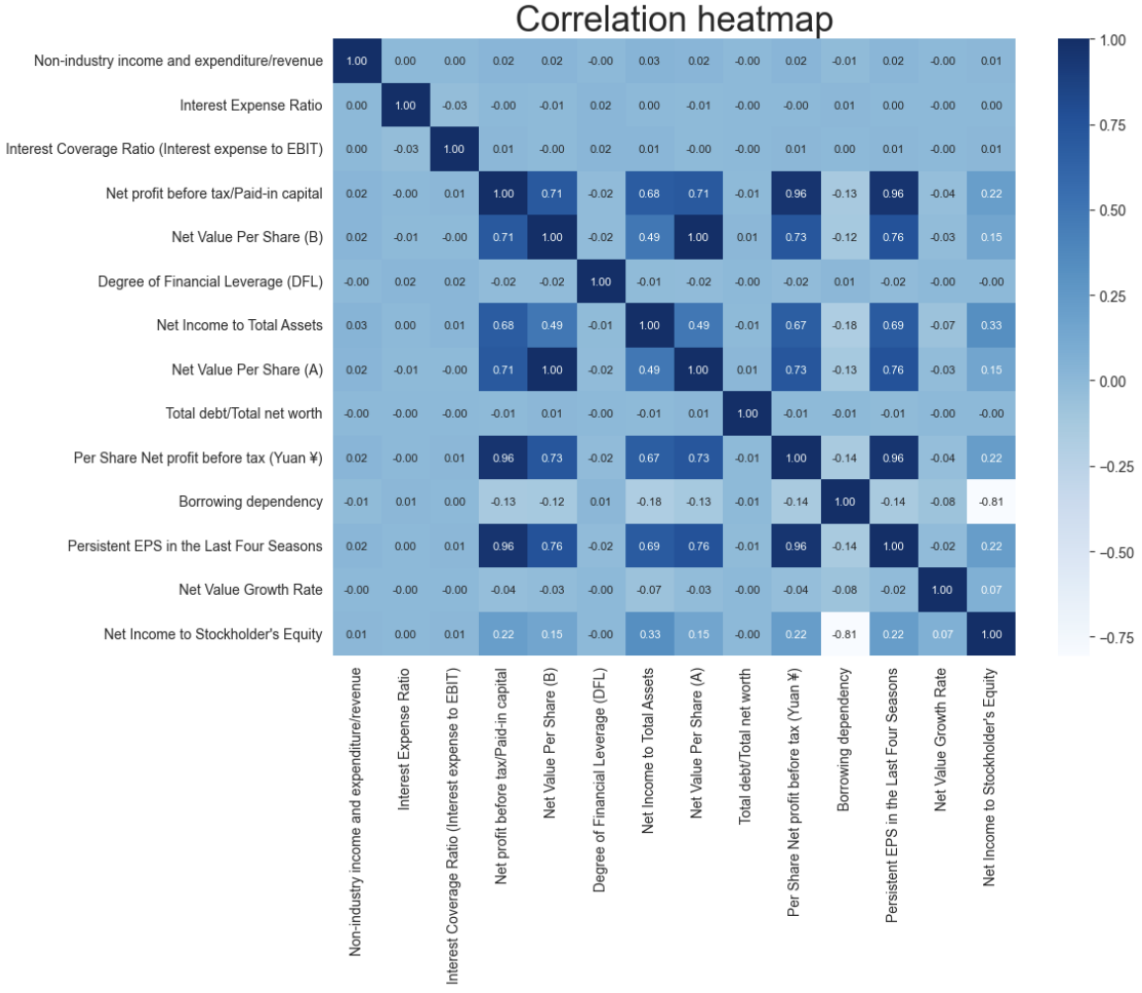

Fig. 2 Correlation Matrix

2.2. Resampling

The research uses two popular resampling methods, namely SMOTE for oversampling and RUS for undersampling. Consequently, the number of samples using oversampling and undersampling are 5279 and 176 respectively.

Before resampling, the full dataset is split into a training set (80%) and a test set (20%) using train_test_split from scikit-learn. Stratified sampling (stratify = y) is employed to preserve the original class distribution in both training and test sets. This ensures the test set to reflect real-world data imbalance, avoiding overestimation of model performance.

SMOTE is an oversampling method that generates synthetic samples for the minority class (bankrupt firms) to balance the class distribution, avoiding the information loss associated with undersampling. For each minority class, SMOTE identifies its k nearest neighbors (k=5) and generates new samples by performing linear interpolation between the sample and its neighbors. SMOTE is applied only to the training set to avoid data leakage.

RUS is an undersampling method that randomly removes samples from the non-bankrupt companies to get a more balanced sample distribution with low computational cost. RUS randomly selects a subset of majority class to ensure that the number of majority samples equals the number of minority samples. Same as SMOTE, RUS was applied exclusively to the training set.

2.3. Model Selection

Five representative classification models are selected, covering linear, non-linear, and tree-based algorithms to ensure diversity. Each model is integrated into a pipeline with StandardScaler, which serves as a normalizer of feature scales, and the corresponding resampling technique (SMOTE/RUS).

LR is a linear model outputting the probability of bankruptcy using the sigmoid function. SVM is a non-linear model that finds the optimal hyperplane to separate classes. The study used kernel = 'rbf' (radial basis function) to detect the non-linear relationships. DT is a tree-based model that partitions the feature space using recursive splitting. RF is an ensemble of decision trees that reduces overfitting through bagging. It is robust to high-dimensional and imbalanced data. XGB is a gradient-boosted tree model that iteratively corrects errors of previous trees.

To further enhance prediction performance, a Soft Voting Classifier is constructed using the top-performing models (RF and XGB). Unlike hard voting, which uses majority class labels, soft voting aggregates class probabilities from base models, assigning higher weights to more confident predictions. To optimize the ensemble model, a Grid Search Cross-Validation (GridSearchCV) is conducted on the training set. The best-performing ensemble model with optimal hyperparameters is finally evaluated on the test set to validate its generalizability.

3. Results

3.1. Metric Selection

This research prioritizes recall and F2-Score as the main metrics, while maintaining AUC-Score for consideration and excluding accuracy, which is the common measure others usually consider.

Recall calculates the proportion of true positive (correct prediction of bankruptcy) over all the corrected predictions of both bankrupt and non-bankrupt firms. In the context of bankruptcy, the significance of recall is more pronounced since the missing of bankruptcy case may result in financial institutions making investment decisions to companies which are actually in bad health, leading to huge financial losses. As a result, recall is one of the most important considerations when comparing model performances.

Precision is calculated as the proportion of correct prediction of bankruptcy over all the bankruptcy cases. It emphasizes the accuracy of positive (bankrupt) prediction.

F2-score takes a comprehensive view by taking both recall and precision into account, with more attention paid to recall. As a result, this research chooses F2-score to incorporate precision as well while providing a more balanced result.

The Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve is a metric that provides a comprehensive evaluation of a classifier's performance. The ROC curve plots the true positive rate (recall) against the false positive rate. The AUC-Score ranges from 0 to 1, where an AUC-Score of 1 represents a classifier that can perfectly distinguish between positive and negative samples at all thresholds, and an AUC-Score of 0.5 indicates a classifier performing no better than random guessing.

Accuracy is a commonly used metric in many classification tasks. However, in the bankruptcy problem, this study chooses not to use it as a primary evaluation metric. The main reason is that accuracy is misleading when the dataset is imbalanced.

3.2. Performance Using Oversampling

Table 1. Model Performance using Oversampling

	Recall for bankruptcy	Precision	F2-Score	AUC-Score
Logistic Regression	0.84	0.18	0.4881	0.9105
Support Vector Machine	0.77	0.15	0.4250	0.9103
Decision Tree	0.43	0.19	0.3417	0.6845
Random Forest	0.68	0.33	0.5639	0.9429
XGBoost	0.61	0.28	0.4927	0.9254

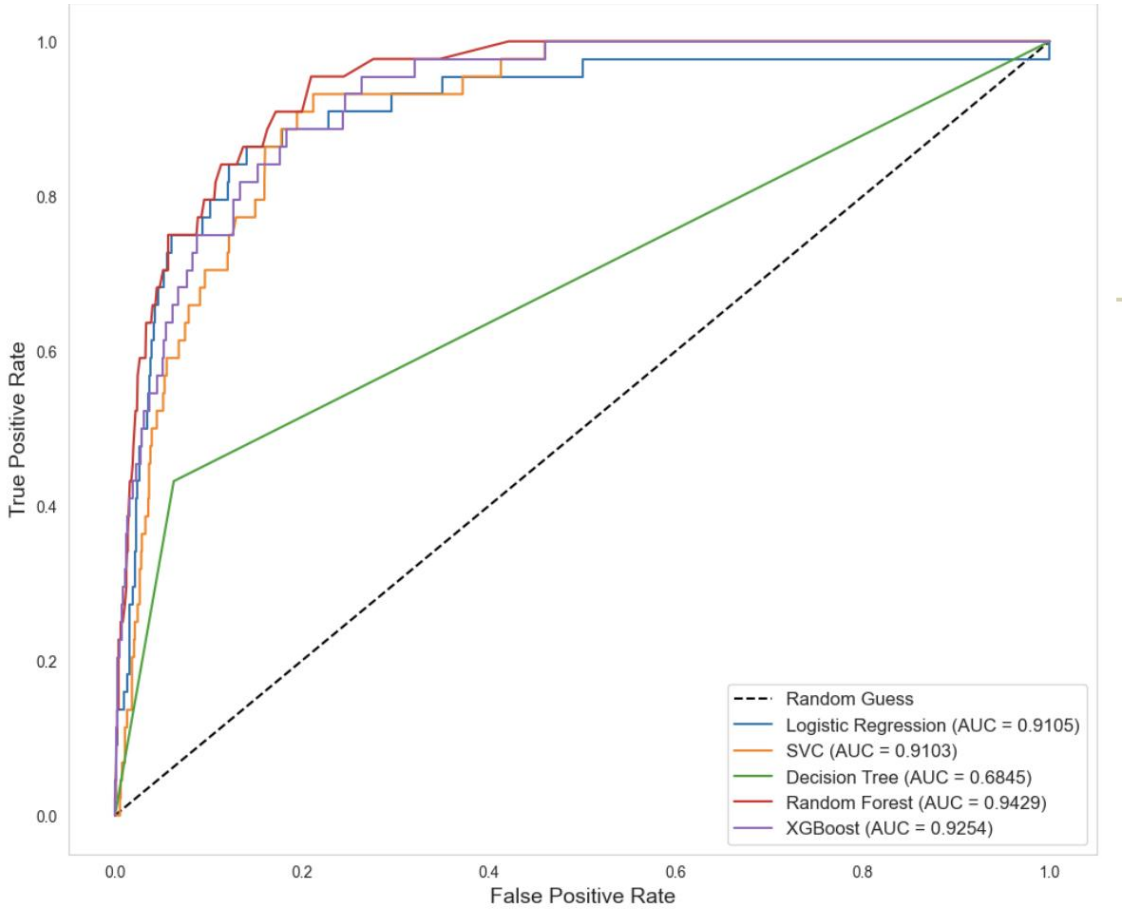

Fig. 3 ROC Curve using Oversampling

As shown in Table 1 and Figure 3, when using oversampling, RF emerges as the most effective model, achieving the highest F2-Score of 0.5639 and a strong AUC-Score of 0.9429. This indicates

a robust balance between a high recall (0.68) and reasonable precision (0.33) for the bankrupt class. XGB also demonstrates a strong and balanced performance, securing the second-highest F2-Score of 0.4927 and AUC-Score of 0.9254.

Notably, while LR achieves the highest recall, this comes at the cost of an extremely low precision of only 0.18. DT shows the weakest performance in this group.

3.3. Performance Using Undersampling

Table 2. Model Performance using Undersampling

	Recall for bankruptcy	Precision	F2-Score	AUC-Score
Logistic Regression	0.80	0.17	0.4642	0.8956
Support Vector Machine	0.84	0.17	0.4637	0.9216
Decision Tree	0.82	0.13	0.4009	0.8193
Random Forest	0.86	0.19	0.5094	0.9296
XGBoost	0.84	0.15	0.4458	0.9066

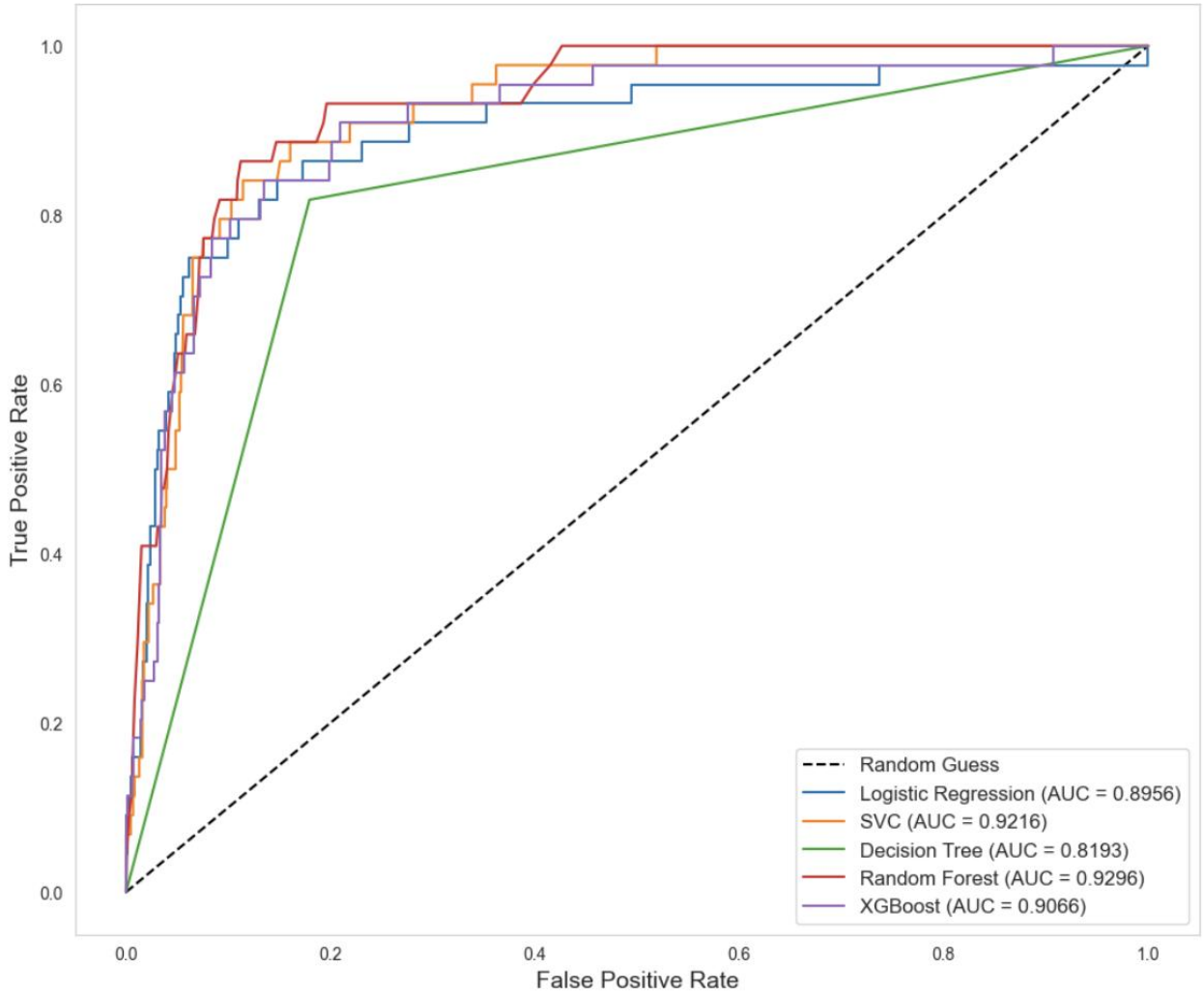

Fig. 4 ROC Curve using Undersampling

As shown in Table 2, when models are trained on the undersampled data, there is a universal trend: all classifiers produce an exceptionally high recall (> 80%). This demonstrates that undersampling makes the models highly aggressive in predicting bankruptcy.

However, this high sensitivity is accompanied by extremely low precision, with all models < 0.19. This indicates that while the models successfully point out most actual bankruptcies, they also

generate a huge number of false positive signals. Despite the above situations, as shown in Table 2 and Figure 4, RF once again achieves the highest F2-Score of 0.5094 and AUC-Score of 0.9296, suggesting that it finds the best trade-off between recall and precision.

3.4. Performance of Voting Classifier

RF and XGB are selected as two base models to form the voting classifier. The ensemble is trained on the oversampled data set using SMOTE and its hyperparameters are optimized using GridSearchCV (rf_max_depth = 20, rf_estimators = 100, xgb_learning rate = 0.1 and xgb estimators = 200). As shown in Table 3, the final tuned model achieves an F2-Score of 0.502, a recall of 0.64, and a precision of 0.27. While its F2-Score does not ultimately surpass the single RF, it provides a stable and robust performance, validating the effectiveness of ensemble methods on this complex task.

Table 3. Performance of Voting Classifier

	Precision	Recall	F1-Score	Support
Not Bankruptcy	0.91	0.94	0.96	1320
Bankruptcy	0.27	0.64	0.38	44
accuracy			0.93	1364
macro avg	0.63	0.79	0.67	1364
weighted avg	0.96	0.93	0.95	1364

4. Discussion

Although the research has showed positive results, the study involves several limitations. The feature selection and resampling methods used are only the fundamental ones, without trying more models like SLR, PLS-DA for feature selection or ADASYN for resampling. Also, the data collected are limited to firm-related financial ratios, while bankruptcy is also influenced by external factors, like interest rates, GDP growth, etc. Also, the analysis is conducted based on companies in Taiwan. Since different countries operate under different accounting standards and economic environments, the results may not be directly applicable to companies in other countries. Thus, future research can proceed, incorporating a wider range of models, considering more general features, and applying the proposed framework to the wider real world.

5. Conclusion

This research addresses the problem of bankruptcy prediction with high feature dimensionality and imbalanced data set. Recognizing the limitations of accuracy, the study prioritizes recall and F2-Score as the main consideration. A systematic comparation of five commonly used machine learning models with a two-layered feature selection method and two resampling techniques is conducted. The findings demonstrate that oversampling provides a more balanced result between recall and precision than undersampling. Among the selected five models, RF and XGB, when paired with SMOTE, are the most robust performers, with a F2-Score of 0.502, a recall of 0.64, and a precision of 0.27. This approach, with a tuned soft-voting classifier, proves to be a reliable and effective way to for financial institutions to handle the complex bankruptcy prediction problem.

References

[1] Daniela R, Mária B, Lucia J. Analysis of the construction industry in the Slovak Republic by bankruptcy model. Procedia - Social and Behavioral Sciences, 2016, 230: 298–306. https://doi.org/10.1016/j.sbspro.2016.09.038

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

[3] Barboza F, Kimura H, Altman E. Machine learning models and bankruptcy prediction. Expert Systems with Applications, 2017, 83: 405–417. https://doi.org/10.1016/j.eswa.2017.04.003

[4] Lu Y, Zeng N, Liu X, Yi S. A new hybrid algorithm for bankruptcy prediction using switching particle swarm optimization and Support Vector Machines. Discrete Dynamics in Nature and Society, 2015, 2015: Article 783262. https://doi.org/10.1155/2015/783262

[5] Härdle W, Lee Y-J, Schäfer D, Yeh Y-R. Variable selection and oversampling in the use of smooth Support Vector Machines for predicting the default risk of companies. Journal of Forecasting, 2009, 28(6): 512–534. https://doi.org/10.1002/for.1103

[6] Son H, Hyun C, Phan D, Hwang H J. Data analytic approach for bankruptcy prediction. Expert Systems with Applications, 2019, 138: 112816. https://doi.org/10.1016/j.eswa.2019.07.033

[7] Liang D, Tsai C F, Wu H T. The effect of feature selection on financial distress prediction. Knowledge-Based Systems, 2015, 73: 289–297. https://doi.org/10.1016/j.knosys.2014.10.010

[8] Chi D-J, Chu C-C. Artificial Intelligence in Corporate Sustainability: Using LSTM and GRU for Going Concern Prediction. Sustainability, 2021, 13(21): 11631. https://doi.org/10.3390/su132111631

[9] Fasano F, Adornetto C, Zahid I, La Rocca M, Montaleone L, Greco G, Cariola A. The dilemma of accuracy in bankruptcy prediction: a new approach using explainable AI techniques to predict corporate crises. European Journal of Innovation Management, 2024, 28(11): 1-22. https://doi.org/10.1108/EJIM-06-2024-0633

[10] D'Ercole A, Me G. A Novel Approach to Company Bankruptcy Prediction Using Convolutional Neural Networks and Generative Adversarial Networks. Machine Learning and Knowledge Extraction, 2025, 7(3): 63. https://doi.org/10.3390/make7030063