

Corporate Bankruptcy Prediction via Mechanics-Grounded Feature Engineering and LightGBM

Shaoke Wang

*Andrew and Erna Viterbi School of Engineering, University of Southern California, Los Angeles,
USA
shaokewa@usc.edu*

Abstract. Predicting corporate bankruptcy is key to managing financial risks. This paper analyzes financial data of Taiwanese listed companies from the Taiwan Economic Journal (1999-2009), and covers the entire prediction process from exploratory analysis to model evaluation. This study considers a dataset of 6,819 instances with 220 positive and 6,599 negative bankruptcy cases, resulting in an imbalanced class of approximately 1:30. Instead of traditional metrics, this paper compute PR-AUC (area under the precision-recall curve), and with no fancy techniques and on a baseline, LightGBM achieved a PR-AUC of 0.487, which is a better result than logistic regression, decision trees and a random forest. The contribution to the field is the introduction of 21 new features based on the bankruptcy risk theory about profitability, pressure from debt, growth, and efficiency and liquidity, which resulted in a PR-AUC of 0.527. The combination of standardized, strict preprocessing, the SMOTE technique for oversampling the minority class, and a strict search for optimal model hyperparameters resulted in a PR-AUC of 0.542 and a ROC-AUC of 0.956. In the final model, the feature based on the bankruptcy risk theory ranked the sixth most important. This order of importance confirms that the introduction of features based on bankruptcy risk theory, rather than the choice of model or resampling techniques, achieved the best results.

Keywords: Bankruptcy prediction, LightGBM, Feature engineering, Class imbalance

1. Introduction

Predicting corporate bankruptcy is part of recognizing financial system risks. The input is the structure of financial statement data, while the output is the prediction for future bankruptcy. This is a critical problem for assessing the price of risk and the efficiency of capital allocation for a credit request or investment decision. Bankruptcy is an event that occurs relatively infrequently, and the sample distribution has an extremely skewed structure. During the learning process, the model is likely to converge toward the majority class, so it is difficult for the model to identify the minority class.

For a long time, the prediction of corporate bankruptcy was based on primitive methods in the construction of risk assessment models using financial indicators. The Z-score model was proposed by Altman and used a creative scoring system based on financial ratios, and for the first time

successfully classified the companies by financial stability level. The Z-score model added impetus to the research of bankruptcy prediction [1]. Most traditional statistical models assume that relationships between variables are linear and are not particularly good at capturing complex, nonlinear relationships between financial indicators. With the emergence of machine learning methods, researchers have begun to apply logistic regression, decision trees, random forests, and gradient boosting models, among others, to corporate bankruptcy predictions [2-5]. Barboza et al. reported that machine learning methods, as compared to traditional statistical methods, provide better solutions to complex financial data, and that random forests and boosting-based methods are particularly superior in predictive ability [6].

Gradient-boosted decision trees dominate the financial risk prediction market. LightGBM has a first-in-class framework and load-balancing structure. It is successfully deployed for a variety of financial predictions and handles high-dimensional structured data with ease [5]. Nevertheless, some of the primary issues of corporate bankruptcy prediction still cannot be addressed by focusing solely on the optimization of the model structure.

First, corporate bankruptcy data usually suffers from severe class imbalance. Since the number of bankrupt companies in the actual market environment is far lower than that of normal operating companies, the model tends to favor the majority class, thereby reducing its ability to identify the minority risk companies. For example, the Taiwan Economic Journal (TEJ) dataset used in this paper contains 6,819 corporate samples, of which only 220 are bankrupt companies, accounting for about 3.23%, forming a typical highly imbalanced classification task. In this scenario, accuracy cannot effectively measure model performance, while the area under the precision-recall curve (PR-AUC) can pay more attention to the identification effect of minority samples and is therefore more suitable as the main evaluation index for bankruptcy prediction tasks [7].

Secondly, existing research typically focuses more on model selection and parameter optimization, while research on how to construct effective feature representations by incorporating knowledge from the financial domain remains insufficient. Corporate financial indicators do not exist in isolation; different indicators reflect different risk factors in the company's operations. For example, profitability affects a company's ability to continue operating, debt levels reflect debt repayment pressure, liquidity indicators reflect short-term funding risks, and operational efficiency indicators reflect resource utilization levels. Therefore, combining and transforming financial indicators according to the company's risk formation mechanisms may be more effective in improving predictive performance than simply increasing model complexity.

Based on the aforementioned research background, this paper proposes a mechanism-grounded feature engineering approach driven by financial mechanisms, and combines it with LightGBM to construct a corporate bankruptcy prediction framework. This paper conducts research based on corporate financial data from the Taiwan Economic News (1999-2009). Building upon the original 95 financial indicators, 21 combined features with financial implications are designed from four dimensions: profitability, debt pressure, growth capacity, and efficiency and liquidity. These features enhance the model's ability to express the structure of corporate financial risk. SMOTE is incorporated to address class imbalance issues, and the relationship between feature structure and model performance is analyzed through comparative experiments. The research aims to uncover the most effective potential assessment indicators for determining whether a company will go bankrupt, providing stronger data support for banks' credit evaluation and investment decisions in practical financial risk management scenarios.

2. Methods

2.1. Dataset

The financial records of the Taiwan Economic Journal for the years 1999 to 2009 (accessible via Kaggle and the UCI Machine Learning Knowledge Base) comprise the dataset. The dataset comprises 6,819 samples and is partitioned into 80/20 hierarchical partitions (training/testing). The training set is made of 5,455 samples, and the testing set is made of 1,364 samples. The class ratios of the two subsets are consistent (the test set is made of 44 bankruptcy samples).

2.2. Models

This paper compares four classifiers: logistic regression, decision tree, random forest, and LightGBM [2-5].

Logistic regression is a linear model used for the binary classification of data. It builds a model of a linear relationship between the features and the categories of the result. Because of the structure of the model, logistic regression is easy to implement and compute, and the predictions are simple to explain. For all of these reasons, it is appropriate to use logistic regression for the analysis of the relationship of variables, such as the prediction of risk in finance. Because of the model's structure, the main drawback of logistic regression is the reliance of the model on linearity. Therefore, the model cannot capture the complexity of the system when it has nonlinear relationships.

A decision tree is a supervised learning model that, as the name suggests, uses a tree-based structure. Like logistic regression, it has the strength of being able to capture the nonlinear relationships of the system and the interactions of the features. However, like the regression model, the tree-based model loses performance if there are fluctuations in the data, because the decision tree is also prone to overfitting.

Random Forest has the advantage of combining several decision trees in the model to improve all the weaknesses of a single decision tree. Random Forest samples data and features to build several decision trees and uses the majority vote of all the trees to predict the result. The Random Forest model has a strength that a single decision tree cannot capture the complexity of the relationships in the data due to the plethora of dimensions and features that it can easily handle, and also has excellent performance in the classification of financial data.

LightGBM is a decision tree model built around a gradient boosting framework. It creates a series of weak learners to iteratively fit residuals of the prior model. From a functionality standpoint, it has advantages over standard gradient boosting. LightGBM has faster training times with greater prediction accuracy on large, high-dimensional data and is great for financial risk prediction of a business.

This paper compares the baseline performance and the effects of feature optimization on four models to derive best practices for predicting corporate bankruptcy.

2.3. Experimental design

Data cleaning consists of three steps. First, a variance filter is fitted on training set 1 using VarianceThreshold (threshold=0) to retain only columns with variance greater than 0 and remove constant columns such as Net Income Flag. Second, the absolute value of the correlation coefficient matrix of the training set features is calculated, and the upper triangle is used to avoid repeated statistics at symmetrical positions. Columns with a correlation coefficient greater than 0.95 with any

other column are removed from both the training and test sets. Table 1 lists representative feature pairs with a correlation coefficient greater than 0.95 on the training set, some of which are completely collinear. Third, the skewness of each feature in the training set is calculated, and the minimum value of the column on both the training and test sets is checked to ensure that it is not less than 0. Only columns that simultaneously meet the conditions of skewness and non-negativity are included and subjected to log1p transformation. After triple filtering of frequent series, variance, and correlation, a total of 18 redundant columns are removed.

Table 1. Redundant feature pairs in training set

Feature 1	Feature 2	ρ
Current Liabilities/Equity	Current Liability to Equity	1.000
Current Liabilities/Liability	Current Liability to Liability	1.000
Debt ratio %	Net worth/Assets	1.000
Operating Gross Margin	Gross Profit to Sales	1.000
Net Value Per Share (A)	Net Value Per Share (C)	0.9998
Operating Gross Margin	Realized Sales Gross Margin	0.9995
Operating Profit Per Share	Operating profit/Paid-in capital	0.9987
Current Liabilities/Equity	Liability to Equity	0.9639

After basic data cleaning, the performance of four models was first compared using the original features without imbalance processing. The PR-AUC metric was used for evaluation. However, due to significant class imbalance in the dataset, with bankrupt companies comprising only 3.23% of the total samples, traditional precision or ROC-AUC metrics are easily affected by the majority class samples and cannot accurately reflect the model's ability to identify the minority of bankrupt companies. Therefore, this paper uses the area under the precision-recall curve as the primary evaluation metric. The results are shown in Table 2. Among the four baseline models, Random Forest and LightGBM achieved better prediction results, with PR-AUCs above 0.45. LightGBM performed best, with a PR-AUC of 0.487.

Table 2. Baseline PR-AUC on the test set

Model	PR-AUC
Logistic Regression	0.037
Decision Tree	0.169
Random Forest	0.457
LightGBM	0.487

Feature engineering is strictly based on 76 features after data cleaning, and four aggregated features: profitability (mean/min/std), debt pressure (mean/max/std), growth (mean/min/std), and efficiency and liquidity (mean/min/std), as detailed in Table 3.

Columns that actually exist in the current feature set are selected from a pre-prepared list of original columns. The mean, extreme values, and standard deviations of these columns are then calculated row-wise to obtain the strength, lower bound, and volatility of each dimension. Based on

this, five cash flow ratio features are created. The central feature, profit-debt balance, is created from the mean debt burden subtracted from the mean profitability. A higher value shows stronger profitability with less leverage, which corresponds to an overall healthier financial profile. The final feature matrix is expanded from 76 columns to 97 columns.

Table 3. The four engineered-feature families (21 features in total)

Family	feats	Representative features
Profitability	4	profitability mean/min/std, cashflow per eps
Debt pressure	6	debt pressure mean/max/std, cashflow to debt pressure, cfo to current liability pressure, profit debt balance
Growth	4	growth mean/min/std, growth debt balance
Efficiency/Liquidity	7	liquidity mean/min/std, liquidity debt balance, cash to current liability pressure, cashflow liquidity balance, cashflow to operating profit

Minority class samples are synthesized using 5-nearest neighbor interpolation, adjusting the training set labels from 176 positive examples/5279 negative examples to approximately 1:1. LightGBM is retrained after each stage, and the PR-AUC of the test set is recorded, resulting in the staged ablation results described in Chapter 3. Since the positive-to-negative sample ratio is approximately 1:30, accuracy is not meaningful; therefore, PR-AUC is used as the primary evaluation metric, with ROC-AUC as a supplementary metric, because the precision-recall curve is more sensitive to minority class identification in severely imbalanced scenarios (Figure 1). Starting with the default parameters of LightGBM, 10 hyperparameters were manually tuned, and the final parameters are shown in Table 4, which balances model capacity, regularization strength, and training cost.

Table 4. Final tuned LightGBM hyperparameters and their roles

Hyperparameter	Value	Role
n estimators	500	Number of boosting rounds
learning rate	0.02	Shrinkage per round
num leaves	35	Max leaves per tree
max depth	20	Max tree depth
min child samples	30	Min samples per leaf
subsample	0.9	Row sampling ratio (bagging)
colsample bytree	0.8	Feature sampling ratio
reg alpha	0.1	L1 regularization
reg lambda	0.5	L2 regularization
min split gain	0.0	Min loss reduction to split

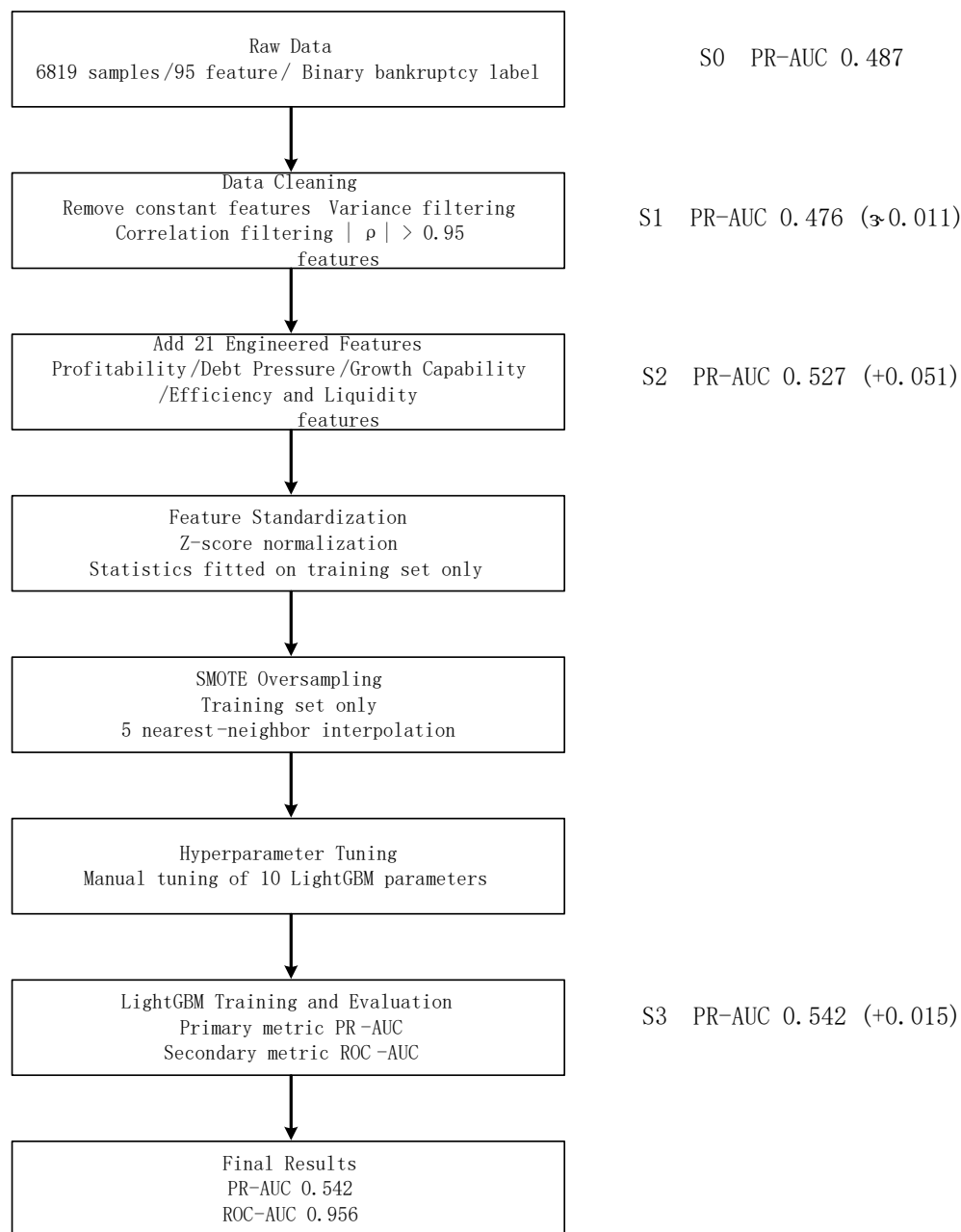


Figure 1. Bankruptcy prediction modeling process (photo/picture credit: original)

3. Results

Table 5 shows the contribution breakdown of each process stage. Data cleaning alone (S1) slightly reduced the PR-AUC from 0.487 to 0.476 because removing 18 collinear columns also discarded some redundant but still valid signals. Adding 21 engineering features (S2) increased the PR-AUC to 0.527, the largest single-step increment in the entire process (+0.051). Standardization, SMOTE, and hyperparameter tuning combined (S3) brought another increment of +0.015, ultimately achieving a PR-AUC of 0.542 and a ROC-AUC of 0.956.

Table 6 shows the top 8 important features in the final model, ranked by split gain. The engineering feature profit debt balance ranked 6th (importance 476), entering the first tier along with

indicators such as inventory turnover rate and expense ratio, indicating that a single manually constructed difference feature can rival the discriminative power of most original ratios.

Table 5. Stage-wise ablation of the LightGBM pipeline

Stage	Processing	# feats	PR-AUC	Δ	ROC-AUC
S0	Raw LightGBM baseline	95	0.487	—	—
S1	+ cleaning	76	0.476	−0.011	—
S2	+ engineered features	97	0.527	+0.051	—
S3	+ scaling + SMOTE + tuning	97	0.542	+0.015	0.956

Table 6. Top-8 features by importance in the final tuned LightGBM

Rank	Feature	Importance
1	Inventory Turnover Rate (times)	640
2	Research and development expense rate	630
3	Operating Expense Rate	543
4	Quick Asset Turnover Rate	512
5	Cash Turnover Rate	488
6	Profit debt balance (engineered)	476
7	Interest-bearing debt interest rate	465
8	Current Asset Turnover Rate	432

4. Discussion

Studies have shown that the engineering feature profit debt balance is the only constructed feature to enter the first tier, indicating that the simple difference between profitability and leverage level has a discriminative power comparable to the strongest original ratio. This supports the core argument of this paper: the feature engineering based on the mechanism drives performance improvement, rather than model selection or resampling itself. This finding is consistent with recent research on gradient boosting ensemble models on imbalanced bankruptcy data. Amirshahi and Lahmiri compared the performance of heterogeneous ensemble models based on XGBoost, LightGBM and CatBoost on highly imbalanced bankruptcy data and found that oversampling may have a negative or limited impact on well-tuned ensemble models [8]. In the staged ablation (Table 5), SMOTE only contributed 0.015 PR-AUC increment, which is far lower than the 0.051 contributed by feature engineering alone. Noh compared the performance of random sampling and sampling strategies based on approximate entropy on small sample Korean bankruptcy data and also emphasized that the choice of resampling method and evaluation threshold, rather than the classifier itself, is the key factor determining the reporting performance [7]. This supports the approach of breaking down each process stage one by one rather than only reporting the final index. Meanwhile, Nazareth and Ramana Reddy's review of the application of machine learning in the financial field points out that the transparent assessment of imbalanced data is still inconsistent in the existing literature, which is also the motivation for this paper to adopt the ablation design based on PR-AUC [9, 10].

5. Conclusion

This paper focuses on the problem of corporate bankruptcy prediction. Based on machine learning methods, it constructs a prediction framework combining financial mechanism-driven feature engineering and the LightGBM model, and systematically studies the nonlinear relationships and class imbalances existing in corporate financial data. The influence of various optimization procedures on the predictive power of corporate bankruptcy is verified through the experimental study of various phases like data preprocessing, feature engineering, construction of the model, and the evaluation of model performance.

Firstly, with respect to the data preprocessing and construction of the model, this paper carries out experiments with the corporate bankruptcy data from the Taiwan Economic Journal. To deal with the redundant information and the highly correlated original financial indicators, data cleansing is conducted by removing constant features, followed by variance and correlation filtering. As for the model to be used, this study compares the prediction powers of four classification models, including logistic regression, decision trees, random forests, and LightGBM. The results of the study show that under the condition of the original financial features, LightGBM postulated a PR-AUC of 0.487 and outperformed the other traditional classification models. This outcome shows that tree-based models that apply ensemble learning have the ability to capture the complex nonlinear correlation of corporate financial indicators more effectively.

Secondly, with regard to the optimization of feature engineering, this study from the perspective of the corporate bankruptcy formation mechanisms, combines four dimensions of profitability, debt pressure, growth capacity, and the liquidity and efficiency of 21 features, and constructs a model. The results of the study show that by introducing mechanism-driven features, the model's PR-AUC improved from 0.476 to 0.527, an increase of 0.051, and that this improvement was the most significant throughout the optimization. This result shows that in order to enhance a model's ability to identify a corporation's potential financial risks, it is more worthwhile to construct features with a reasonable representation of the domain through expertise, as opposed to increasing the complexity of the model.

Finally, regarding model optimization and comprehensive evaluation, this paper implements the SMOTE method to address the class imbalance problem caused by the small number of bankruptcy samples and optimizes the hyperparameters of the LightGBM model. The final model achieves prediction results of PR-AUC 0.542 and ROC-AUC 0.956, representing a significant improvement compared to the baseline model. Ablation experiments further validate the actual contributions of different optimization modules, with feature engineering having the most significant impact on model performance improvement, while data sampling and parameter tuning further enhance the model's generalization ability.

In summary, this paper demonstrates that structured feature engineering based on financial mechanism design can effectively improve the performance of corporate bankruptcy prediction models, and combines it with LightGBM to construct an effective method suitable for predicting highly imbalanced financial risks. This research not only provides a feasible technical solution for corporate credit risk assessment and risk early warning, but also illustrates the significant value of combining machine learning models with knowledge from the financial domain. Future research can further combine time-series financial data, cross-market corporate samples, and interpretable machine learning methods (such as SHAP) to further enhance the model's generalization ability and decision interpretability.

References

- [1] Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4), 589–609. <https://doi.org/10.1111/j.1540-6261.1968.tb00843.x>
- [2] Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2), 215–242.
- [3] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth International Group.
- [4] Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- [5] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 3146–3154.
- [6] Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, 83, 405–417. <https://doi.org/10.1016/j.eswa.2017.04.006>
- [7] Noh, S. H. (2023). Comparing the performance of corporate bankruptcy prediction models based on imbalanced financial data. *Sustainability*, 15(6), 4721. <https://doi.org/10.3390/su15064721>
- [8] Amirshahi, B., & Lahmiri, S. (2024). Bankruptcy prediction using optimal ensemble models under balanced and imbalanced data. *Expert Systems*, 41(8), e13599. <https://doi.org/10.1111/exsy.13599>
- [9] Nazareth, N., & Ramana Reddy, Y. V. (2023). Financial applications of machine learning: A literature review. *Expert Systems with Applications*, 219, 119640. <https://doi.org/10.1016/j.eswa.2023.119640>
- [10] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.