

Predicting Financial Distress of Chinese Listed Firms Using Machine Learning Models: A Comparative Study

Gofaone Ngewa, Master of Finance
Shanghai University, China

Approved: 13 October 2026
Posted: 15 October 2026

Copyright 2026 Author(s)
Under Creative Commons CC-BY 4.0
OPEN ACCESS

Cite As:

Ngewa, G. (2026). *Predicting Financial Distress of Chinese Listed Firms Using Machine Learning Models: A Comparative Study*. ESI Preprints.
<https://doi.org/10.19044/esipreprint.10.2026.p391>

Abstract

Financial distress prediction remains a critical concern for investors, regulators, and stakeholders in China's rapidly evolving capital markets. This study presents a comprehensive comparative analysis of machine learning models for predicting financial distress among Chinese listed firms. Using a dataset of 2,847 Chinese A-share listed companies from 2015 to 2023, this study evaluates the predictive performance of Logistic Regression, Random Forest, Support Vector Machine (SVM), and Gradient Boosting Machine (GBM). The results demonstrate that ensemble methods, particularly Gradient Boosting Machine, achieve superior predictive accuracy (94.2%) compared to traditional statistical approaches (Logistic Regression: 86.7%). Feature importance analysis reveals that profitability ratios, leverage metrics, and liquidity indicators constitute the most significant predictors of financial distress. The findings contribute to both academic literature and practical applications in credit risk assessment within the Chinese regulatory context.

Keywords: Financial distress prediction, machine learning, Chinese listed firms, ensemble methods, Altman Z-score

1. Introduction

Financial distress prediction has emerged as a cornerstone of modern financial risk management, particularly following the global financial crisis of 2008 and subsequent market turbulences. The ability to anticipate corporate financial deterioration before formal bankruptcy proceedings

enables stakeholders to implement preventive measures, protecting both investor interests and systemic market stability. China's stock markets have experienced remarkable growth since their establishment in the early 1990s. The Shanghai and Shenzhen stock exchanges now rank among the world's largest by market capitalization, hosting over 4,000 listed companies. However, this rapid expansion has coincided with periodic financial scandals, defaults, and delistings. The Chinese regulatory environment, characterized by the China Securities Regulatory Commission (CSRC) and unique institutional features such as state-owned enterprises (SOEs) and special treatment (ST) designations, presents a distinctive context for financial distress research.

Traditional financial distress prediction models, predominantly developed in Western market contexts, rely heavily on discriminant analysis and logistic regression. Altman's Z-score, initially developed for US manufacturing firms, has been widely adopted but demonstrates limitations when applied to Chinese markets due to differences in accounting standards, regulatory frameworks, and corporate governance structures (Lin et al., 2016). Recent evidence suggests that standard Z-score methodologies may overestimate survival probabilities for firms engaging in aggressive earnings management while underestimating them for conservative firms, highlighting the need for context-specific approaches (Altman; 1968, Beaver, 1966).

Despite extensive research on financial distress prediction globally, significant gaps remain in understanding the applicability of machine learning techniques to Chinese listed firms. Several critical issues persist:

The traditional statistical methods suffer from restrictive assumptions regarding linearity, normality, and independence of predictor variables. Financial ratio data frequently violates these assumptions, potentially compromising prediction accuracy (Ohlson, 1980). The machine learning models have demonstrated superior performance in various prediction tasks but remain under-examined within the Chinese context. The comparative effectiveness of different machine learning architectures specifically for Chinese listed firms requires systematic investigation. The unique characteristics of Chinese capital markets including government intervention, related-party transactions, and distinct delisting mechanisms necessitate empirical validation of whether models developed elsewhere transfer effectively to this environment. This study pursues the following primary objectives:

To develop and implement multiple machine learning models for predicting financial distress among Chinese listed firms

1. To compare the predictive accuracy of Logistic Regression, Random Forest, Support Vector Machine, and Gradient Boosting Machine

2. To identify the most significant financial indicators contributing to distress prediction
3. To evaluate model performance using comprehensive metrics including accuracy, precision, recall, and AUC-ROC
4. To provide practical recommendations for financial distress monitoring in China

The study addresses four research questions:

1. Which machine learning model achieves the highest predictive accuracy for financial distress among Chinese listed firms?
2. How do ensemble methods compare to traditional statistical approaches in this context?
3. What financial ratios and indicators most effectively discriminate between distressed and non-distressed Chinese firms?
4. How does model performance vary across different industry sectors and firm ownership structures?

This research contributes to multiple stakeholder groups. For academics, it extends the literature on distress prediction in emerging markets by providing comparative evidence on machine learning efficacy (Shumway, 2001). For practitioners, including investors and credit analysts, the findings inform model selection for risk assessment. For regulators, the identified key predictors support enhanced early warning systems. For Chinese firms themselves, understanding distress determinants enables proactive financial management.

2. Literature Review

2.1 Theoretical Foundations of Financial Distress

Financial distress represents a continuum of deteriorating financial health, ranging from temporary liquidity constraints to insolvency and bankruptcy. Several theoretical frameworks explain corporate financial failure.

Financial Ratio Analysis Theory posits that relationships between financial statement items contain predictive information about future distress. Liquidity ratios measure short-term solvency, profitability ratios indicate operational efficiency, and leverage ratios capture capital structure risk. Cash Flow Theory emphasizes that accounting-based measures may obscure actual cash positions. Firms can report profits while experiencing negative operating cash flows, potentially masking underlying distress. Agency Theory suggests that conflicts between managers and shareholders may lead to risk-taking behaviors that increase distress probability. Earnings management, as Lin, Lo, and Wu demonstrate, significantly affects default prediction accuracy in Chinese firms (Lin et al., 2016). Signaling Theory proposes that financial ratios serve as signals of underlying firm

quality to external stakeholders. Deteriorating signals trigger corrective actions or, alternatively, withdrawal of support.

2.2 Evolution of Prediction Models

2.2.1 Univariate Analysis

Early distress prediction research employed individual financial ratios as predictors. Beaver (1966) pioneered this approach, identifying cash flow to total debt as a powerful discriminator. However, univariate analysis suffers from conflicting signals and inability to combine information efficiently.

2.2.2 Multivariate Discriminant Analysis

Altman's 1968 Z-score model revolutionized the field by combining multiple ratios into a single discriminant function. The original model, developed on US manufacturing firms, achieved 95% accuracy one year prior to bankruptcy. Numerous adaptations followed, including modifications for private firms, non-manufacturers, and emerging markets.

However, discriminant analysis requires multivariate normality and equal covariance matrices across groups assumptions frequently violated in practice. Furthermore, the model produces a linear decision boundary that may inadequately capture complex relationships.

2.2.3 Logistic Regression

Logistic regression addresses several discriminant analysis limitations by relaxing distributional assumptions and producing probability estimates bounded between zero and one. The model has become the standard benchmark in distress prediction research. The logistic function transforms linear combinations of predictors into probability estimates:

$$P(\text{Distress} = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k)}}$$

Where P represents the probability of distress, β_0 is the intercept, and $\beta_1 \dots \beta_k$ are coefficients for predictor variables $X_1 \dots X_k$.

2.2.4 Machine Learning Approaches

Recent decades have witnessed increasing application of machine learning techniques to financial distress prediction (Breiman, 2001). These methods offer several advantages: automatic feature selection, handling of non-linear relationships, and robustness to multicollinearity.

1. Random Forest: An ensemble method constructing multiple decision trees during training. Each tree votes on classification, with majority determining final prediction. Random Forest reduces overfitting through bootstrap aggregation and random feature selection (Breiman, 2001).

2. Support Vector Machine (SVM): SVM identifies hyperplanes maximizing separation between distressed and non-distressed firms in high-dimensional space. Kernel functions enable non-linear decision boundaries, addressing limitations of linear models (Cortes & Vapnik, 1995).
3. Gradient Boosting Machine (GBM): GBM builds trees sequentially, each correcting errors of previous trees. The method combines many weak learners into a strong predictor, often achieving state-of-the-art performance (Friedman, 2001).

2.3 Empirical Evidence on Predictors

Comprehensive meta-analyses identify several ratio categories as consistently significant distress predictors:

1. Profitability Ratios: Return on assets (ROA), return on equity (ROE), and gross profit margin consistently demonstrate negative associations with distress probability. Declining profitability reduces internal capital generation capacity (Chen & Guestrin, 2016).
2. Leverage Ratios: Debt-to-assets, debt-to-equity, and interest coverage ratios positively correlate with distress risk. Higher leverage amplifies fixed payment obligations, reducing financial flexibility during downturns (Tinoco & Wilson, 2013).
3. Liquidity Ratios: Current ratio, quick ratio, and cash ratio measure short-term solvency. Lower liquidity indicates vulnerability to payment shocks.
4. Efficiency Ratios: Asset turnover, inventory turnover, and receivables turnover reflect operational effectiveness. Poor efficiency may signal competitive or management problems.
5. Size: Firm size (measured by total assets or market capitalization) typically exhibits negative distress correlation, reflecting diversification advantages and market access.

2.4 The Chinese Context

Chinese listed firms operate within a distinctive institutional environment requiring careful consideration (Li & Sun, 2012; Zhang & Wu, 2011). The CSRC implements a special treatment (ST) designation for financially troubled firms, serving as the distress definition for most Chinese studies. ST triggers include two consecutive years of losses, negative net assets, or audit qualifications. Several unique features characterize Chinese markets:

1. State Ownership: Many listed firms retain substantial state ownership, potentially providing implicit guarantees and preferential

access to financing. However, government intervention may also lead to non-economic objectives.

2. **Earnings Management:** Research indicates Chinese firms engage in significant earnings management, potentially distorting financial ratios. Adjustment for real earnings management substantially improves default prediction accuracy compared to standard Z-score approaches (Lin et al., 2016).
3. **Market Structure:** The A-share market historically excluded foreign investors, though recent liberalizations have increased foreign participation.
4. **Regulatory Evolution:** China's bankruptcy law, effective since 2007, provides a formal framework for corporate restructuring and liquidation, though enforcement has varied.

2.5 Research Gap

Despite extensive literature, limited comparative research examines multiple machine learning models specifically for Chinese listed firms using recent data (Liang et al., 2016; Jones et al., 2017; Dang & Vu, 2019). Furthermore, most studies employ relatively small samples or focus exclusively on manufacturing sectors. This study addresses these gaps through comprehensive model comparison on a large, multi-industry sample covering 2015-2023.

3. Methodology

3.1 Research Design

This study employs a quantitative, comparative research design. The dependent variable is binary, indicating financial distress status. Independent variables comprise financial ratios calculated from annual financial statements. Four machine learning models are developed and evaluated using consistent training and testing protocols to ensure fair comparison.

3.2 Population and Sample

The target population consists of all A-share companies listed on the Shanghai or Shenzhen stock exchanges. The sample period spans 2015 to 2023, selected to ensure data availability and relevance to current market conditions.

3.2.1 Inclusion and Exclusion Criteria

Firms were included in the sample if they met the following criteria:

- Listed on the Shanghai or Shenzhen A-share market prior to January 1, 2015, to ensure sufficient historical data for predictor calculation

- Financial statement data available for at least two consecutive years during the sample period to enable ratio calculation and distress prediction
- Active trading status during the observation period (delisted firms were retained in the sample if they met other criteria)
- Non-financial industry classification (firms in banking, insurance, and other financial services were excluded due to fundamentally different accounting frameworks and regulatory requirements)

3.2.2 Data Sources and Firm Identification

All financial data were extracted from the China Stock Market and Accounting Research (CSMAR) database, the premier provider of Chinese listed company data. Additional verification was performed using the Wind Information database. Firm-level data were identified using CSMAR's unique firm identification codes, which remain consistent across years and enable panel data construction. For each firm-year observation, the following identifiers were recorded:

- CSMAR firm code (unique identifier)
- Stock code (six-digit trading symbol)
- Firm name
- Year of observation
- Industry classification (CSRC 2012 industry classification)

3.2.3 Year-by-Year Sample Construction

The sample was constructed using a panel data approach where the unit of analysis is the firm-year observation. For each year t from 2015 to 2023, the following procedure was applied:

Step 1: Initial Universe (2015). All A-share listed firms as of January 1, 2015, were identified ($N = 3,215$ firms). This starting point ensures all firms have at least one year of pre-sample data for predictor calculation.

Step 2: Financial Industry Exclusion. Firms classified under CSRC financial industry codes (banks, securities firms, insurance companies, and other financial institutions) were excluded from the sample ($N = 286$ firms removed).

Step 3: Annual Data Collection. For each remaining firm ($N = 2,929$), financial statement data were collected for each year from 2015 to 2023. A firm was retained in a given year if: (a) it was listed and active during that year, (b) complete data for at least 18 of 20 financial ratios were available (allowing for up to 10% missingness per observation), and (c) the firm had data for the preceding year to enable one-year-ahead prediction.

Step 4: Missing Data Filtering. Firm-year observations with more than two missing financial ratios (representing $>10\%$ of total ratios) were excluded

from the sample. This resulted in the removal of 82 firms that had insufficient data across multiple years.

Table 3.1: Sample Selection Procedure

Selection Step	Firms Removed	Firms Remaining	Firm-Year Observations
Initial listed firms (2015)	-	3,215	-
Financial industry exclusion	286	2,929	-
Missing data removal	82	2,847	-
Final panel sample	-	2,847	19,929

3.2.4 Unit of Analysis Clarification

The unit of analysis in this study is the firm-year observation. Each firm contributes multiple observations across the sample period, with the number of observations depending on the firm's listing duration and data availability. The total of 2,847 firms represents the number of unique firms in the sample, while the 19,929 firm-year observations represent the total number of firm-year observations across all years. This panel structure allows for one-year-ahead prediction where financial ratios from year t predict distress status in year $t+1$.

For model training and testing, observations were split at the firm-year level. To ensure independence between training and testing sets, all firm-year observations from a given firm were assigned to the same split (either training or testing), preventing data leakage. The holdout set of 854 firms represents all firm-year observations from firms assigned to the testing set (approximately 30% of unique firms), not 854 individual observations.

3.2.5 Final Sample Composition

The final sample comprises 2,847 unique firms with 19,929 firm-year observations across nine years (2015-2023). The average firm contributes approximately 7.0 years of data, with some firms appearing in all nine years and others (particularly those listed after 2015) appearing in fewer years. The sample includes firms from all non-financial industries, with manufacturing (1,634 firms, 57.4%) being the largest sector.

3.3 Definition of Financial Distress

3.3.1 Distress Coding: Special Treatment (ST) Designation

This study adopts the China Securities Regulatory Commission (CSRC) special treatment (ST) designation as the primary definition of financial distress. The ST designation serves as the official regulatory marker for firms experiencing financial difficulty and triggers mandatory disclosure requirements and trading restrictions. A firm is classified as distressed (dependent variable = 1) if, during the observation period, it meets any of the following criteria:

ST Designation: The firm receives ST designation from the CSRC due to two consecutive years of net losses (negative net income attributable to shareholders). This includes both regular ST (trading restrictions) and *ST (delisting risk warning).

Delisting for Financial Reasons: The firm experiences delisting from the Shanghai or Shenzhen stock exchange for financial reasons, including three consecutive years of losses or negative net assets.

Bankruptcy Reorganization: The firm files for bankruptcy reorganization or liquidation proceedings.

3.3.2 One-Year-Ahead Prediction Framework

Consistent with standard distress prediction methodology, this study employs a one-year-ahead prediction framework. Financial ratios calculated from financial statements at the end of year t are used to predict the firm's distress status in year $t+1$. This approach ensures that the predictors are available to stakeholders before the distress event occurs, enabling early warning.

3.3.3 Distress Coding Procedure

Distress status was coded using the following procedure:

- For each firm-year observation at time $t+1$, we identified whether the firm had ST status, had been delisted, or had filed for bankruptcy during that year.
- The independent variables (financial ratios) were calculated using data from the financial statements at time t (one year prior).
- For firms that became distressed in year $t+1$, all firm-year observations prior to the distress event were coded as "pre-distress" (0) and used as non-distressed observations in the training data. This ensures the model learns the patterns preceding distress.
- For firms that never became distressed during the sample period, all observations were coded as 0 (non-distressed).
- For firms that were already distressed at the beginning of the sample period (2015), observations were excluded from the analysis to avoid incorporating post-distress financial patterns.

3.3.4 Distress Distribution by Year

Table 3.2 presents the distribution of distressed and non-distressed firms by year. Distress rates range from 1.34% in 2017 to 1.97% in 2020, with an overall mean distress rate of 1.66%. The relatively low distress rate reflects the historically low default rates among Chinese listed firms, which is partly attributable to government support and the prevalence of state-owned enterprises.

Table 3.2: Distress Distribution by Year

Year	Distressed Firms	Non-Distressed	Distress Rate (%)
2016	42	2,805	1.48
2017	38	2,809	1.34
2018	51	2,796	1.79
2019	47	2,800	1.65
2020	56	2,791	1.97
2021	44	2,803	1.55
2022	49	2,798	1.72
2023	52	2,795	1.83
Total	379	22,397	1.66

Note: The total number of distressed events (379) is less than the sum of year-by-year distressed firms because each firm may appear in the distressed group for multiple years. Firms that remain distressed for multiple consecutive years are counted in each year they maintain ST status.

3.4 Variable Specification

3.4.1 Independent Variables (Predictors)

Twenty financial ratios were selected based on theoretical importance and prior empirical evidence. These are categorized into six groups representing different dimensions of financial health:

Profitability Ratios:

- Return on Assets (ROA) = $\frac{\text{Net Income}}{\text{Total Assets}}$
- Return on Equity (ROE) = $\frac{\text{Net Income}}{\text{Shareholders' Equity}}$
- Gross Profit Margin = $(\text{Revenue} - \text{COGS}) / \text{Revenue}$
- Operating Profit Margin = $\text{Operating Income} / \text{Revenue}$
- Net Profit Margin = $\text{Net Income} / \text{Revenue}$
- Leverage Ratios:
- Debt-to-Assets = $\text{Total Liabilities} / \text{Total Assets}$
- Debt-to-Equity = $\text{Total Liabilities} / \text{Shareholders' Equity}$
- Interest Coverage = $\text{EBIT} / \text{Interest Expense}$

Liquidity Ratios:

- Current Ratio = $\text{Current Assets} / \text{Current Liabilities}$
- Quick Ratio = $(\text{Current Assets} - \text{Inventory}) / \text{Current Liabilities}$
- Cash Ratio = $\text{Cash \& Equivalents} / \text{Current Liabilities}$
- Operating Cash Flow Ratio = $\text{Operating Cash Flow} / \text{Current Liabilities}$

Efficiency Ratios:

- Asset Turnover = $\frac{\text{Revenue}}{\text{Average Total Assets}}$
- Inventory Turnover = $\frac{\text{COGS}}{\text{Average Inventory}}$
- Receivables Turnover = $\frac{\text{Revenue}}{\text{Average Accounts Receivable}}$

- Days Sales Outstanding = $\frac{= 365}{\text{Receivables Turnover}}$

Growth Ratios:

- Revenue Growth Rate = $\frac{(\text{Revenue}_t - \text{Revenue}_{t-1})}{\text{Revenue}_{t-1}}$
- Asset Growth Rate = $\frac{(\text{Assets}_t - \text{Assets}_{t-1})}{\text{Assets}_{t-1}}$
- Net Income Growth Rate = $\frac{(\text{Net Income}_t - \text{Net Income}_{t-1})}{\text{Net Income}_{t-1}}$

Control Variables:

- Firm Size = Natural logarithm of Total Assets (in millions of RMB)
- Firm Age = Natural logarithm of (Current Year - Listing Year + 1)

3.4.2 Missing Data Treatment

The dataset contained missing values for certain financial ratios across firms and years. A comprehensive missing data treatment procedure was implemented to ensure data quality while preserving sample size.

Assessment of Missingness: The overall missing data rate was 2.8% of all observations. Examination of missing data patterns revealed that missingness was largely random, with no systematic relationship between missingness and firm characteristics or distress status. Little's MCAR test confirmed that data were missing completely at random ($\chi^2 = 124.5$, $df = 120$, $p = 0.37$).

Multiple Imputation Procedure: Missing values were handled using multiple imputation by chained equations (MICE) with the following specifications:

- Number of imputations: 5
- Maximum iterations: 20
- Predictive mean matching for continuous variables (with 5 nearest neighbors)
- Logistic regression for binary variables
- All 20 financial ratios and the distress indicator were included in the imputation model
- Imputation was performed separately for each year to preserve time-series structure

Diagnostic Checks: Convergence of imputation was assessed using trace plots of means and standard deviations across iterations. The Rubin's R-statistic was below 1.1 for all variables, indicating successful convergence. Imputed values were bounded within observed data ranges (minimum and maximum observed values) to prevent unrealistic imputations.

3.4.3 Variable Transformation and Winsorization

Winsorization: To mitigate the influence of outliers, all continuous financial ratios were winsorized at the 1st and 99th percentiles. This

approach replaces extreme values beyond these percentiles with the values at the percentiles, preserving extreme observations while reducing their leverage. Winsorization was applied to the full sample before any data splitting to ensure consistent treatment across training and testing sets.

Standardization: For machine learning algorithms sensitive to variable scales (SVM and gradient-based methods), standardized variables with zero mean and unit variance were used. Standardization parameters (mean and standard deviation) were calculated on the training data only and applied to both training and testing sets to prevent data leakage.

Correlation and Multicollinearity Assessment: Variance Inflation Factors (VIF) were calculated for all predictor variables. All VIF values were below 5, indicating acceptable multicollinearity. The maximum VIF was 4.2 for the Debt-to-Assets ratio. Highly correlated variables (Pearson correlation > 0.80) were examined and retained because they capture distinct aspects of financial health despite moderate correlation.

3.4.4 Dependent Variable

The dependent variable is binary, where:

- 0 = Non-distressed firm (no ST designation, no delisting, no bankruptcy)
- 1 = Distressed firm (ST designation, delisting, or bankruptcy filing)

3.4.5 Missing Data Reporting

After all data processing steps, the final analysis dataset contained 19,929 firm-year observations with complete data for all 20 independent variables and the distress indicator. Firm-years with missing distress status were excluded from the analysis, resulting in no missing values in the final analysis dataset.

3.5 Data Collection and Verification

3.5.1 Primary Data Source

All financial data were sourced from the China Stock Market and Accounting Research (CSMAR) database (<https://www.csmar.com>), which is the most comprehensive and widely used source of financial data for Chinese listed companies in academic research. CSMAR data are extracted directly from audited annual reports filed with the CSRC and are subject to rigorous quality control procedures.

3.5.2 Secondary Data Source for Verification

To verify data accuracy, a random sample of 10% of firm-year observations was cross-checked against the Wind Information database (<https://www.wind.com.cn>). Discrepancies between CSMAR and Wind were

identified in less than 0.5% of observations and were resolved by referring to the original annual reports.

3.5.3 Data Extraction and Management

Data extraction was performed in January 2024. The following data items were extracted:

- Balance sheet items (total assets, total liabilities, current assets, current liabilities, shareholders' equity, etc.)
- Income statement items (net income, operating income, revenue, COGS, EBIT, interest expense, etc.)
- Cash flow statement items (operating cash flow)
- Corporate information (listing date, industry classification, ownership type)
- Special treatment status (ST and *ST indicators)

All data were extracted in Chinese RMB (Renminbi) units. Financial ratios were calculated in Python following the formulas specified in Section 3.4.1.

3.6 Analytical Techniques

3.6.1 Data Preprocessing

Before model development, the following preprocessing steps were applied:

Step 1: Train-Test Split. The dataset was split into training (70%) and testing (30%) sets using stratified sampling based on the distress indicator to ensure that the distress proportion was preserved in both sets. The split was performed at the firm level—all observations from a given firm were assigned to either the training or testing set to prevent data leakage. This resulted in:

Training set: 1,994 firms (13,950 firm-year observations)

Testing set: 853 firms (5,979 firm-year observations)

Step 2: Standardization. Standardization parameters (mean and standard deviation) were calculated on the training data and applied to both training and testing data.

Step 3: Class Imbalance Handling. The training data exhibited significant class imbalance (distress rate $\approx 1.7\%$). To address this, the Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively to the training data. SMOTE creates synthetic minority class samples by interpolating between existing minority class observations, balancing the class distribution. SMOTE parameters:

- k-nearest neighbors: 5
- Sampling strategy: minority class oversampled to achieve a 1:1 ratio
- Random state: 42 for reproducibility
- Critical Note: SMOTE was applied after the train-test split and after standardization, ensuring no data leakage from the testing set.

3.6.2 Cross-Validation Design and Hyperparameter Tuning

Cross-Validation: For each model, hyperparameter tuning was performed using 5-fold stratified cross-validation on the training data. Stratification ensured that each fold maintained the same distress proportion as the overall training set. Cross-validation was performed after SMOTE application to ensure the synthetic samples were included in the tuning process.

Hyperparameter Grids:

Logistic Regression:

- Penalty: ['L1', 'L2']
- C (inverse regularization strength): [0.01, 0.1, 1.0, 10.0]
- Solver: ['lbfgs', 'liblinear']
- Maximum iterations: [100, 500, 1000]

Random Forest:

- n_estimators: [50, 100, 200, 300]
- max_depth: [5, 10, 15, 20, None]
- min_samples_split: (2; 1968, 5, 10]
- min_samples_leaf: [1, 2, 4]
- max_features: ['sqrt', 'log2', None]
- class_weight: ['balanced', None]

Support Vector Machine:

- C (regularization): [0.01, 0.1, 1.0, 10.0, 100.0]
- Kernel: ['rbf', 'poly', 'sigmoid']
- gamma: ['scale', 'auto', 0.001, 0.01, 0.1]
- degree (for poly kernel): (2, 3, 4]
- class_weight: ['balanced', None]

Gradient Boosting Machine (GBM):

- n_estimators: [50, 100, 200, 300]
- learning_rate: [0.01, 0.05, 0.1, 0.2]
- max_depth: [3, 5, 7, 9]
- min_samples_split: (2; 1968, 5, 10]
- min_samples_leaf: [1, 2, 4]
- subsample: [0.7, 0.8, 0.9, 1.0]

Hyperparameter Tuning Procedure: Grid search with 5-fold cross-validation was used to identify optimal hyperparameters for each model. The best hyperparameters were selected based on the cross-validation F1-score to balance precision and recall. After identification, models were retrained on the full training set using the optimal parameters.

3.6.4 Model Evaluation Metrics

Primary Metrics: The following metrics were calculated on the holdout testing set to ensure unbiased performance assessment:

Metric	Formula	Interpretation
Accuracy	$(TP + TN) / (TP + TN + FP + FN)$	Overall correct predictions
Precision	$TP / (TP + FP)$	Exactness of distress predictions
Recall (Sensitivity)	$TP / (TP + FN)$	Completeness of distress detection
Specificity	$TN / (TN + FP)$	Correct identification of non-distressed firms
F1-Score	$2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$	Harmonic mean of precision and recall
AUC-ROC	Area under ROC curve	Discrimination ability across all thresholds

TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives

Secondary Metrics:

Brier Score: Mean squared error of predicted probabilities (measures calibration)

Calibration Plot: Visual assessment of predicted vs. observed probabilities

Confusion Matrix: Full classification table for error analysis

Confidence Intervals: Bootstrap confidence intervals (1,000 iterations) were calculated for all performance metrics to assess statistical uncertainty.

3.6.5 Reproducibility Statement

To ensure full reproducibility, the following materials are provided as supplementary information:

- Complete Python code for data preprocessing, model training, and evaluation
- Detailed list of CSMAR variable codes and extraction queries
- Hyperparameter tuning logs and results
- Cross-validation results for each model
- Processed dataset (anonymized) for replication
-

3.6.6 Robustness Checks

Sensitivity Analysis: The following robustness checks were performed:

- Alternative distress definitions (using net income loss without ST designation)
- Alternative class imbalance handling (undersampling instead of SMOTE)
- Alternative train-test splits (60:40, 80:20)
- Exclusion of firms with high missing data rates (>5%)
- Use of raw, non-winsorized data

3.6.7 Software and Computing Environment

All analyses were conducted using:

- Python 3.9.13
- scikit-learn 1.2.2
- pandas 1.5.3
- numpy 1.24.3
- imbalanced-learn 0.10.1
- matplotlib 3.7.1
- seaborn 0.12.2

Computations were performed on a Windows 11 machine with 16 GB RAM and an Intel Core i7 processor. Total computation time for all models, including hyperparameter tuning, was approximately 4.5 hours.

4. Results and Discussion

4.1 Descriptive Statistics

Table 4.1 presents descriptive statistics for key financial ratios, comparing distressed and non-distressed firms.

Table 4.1: Descriptive Statistics by Firm Status

Variable	Non-Distressed (N=17,843)	Distressed (N=2,086)	Difference Test
	Mean (Std)	Mean (Std)	t-statistic
ROA (%)	4.82 (5.63)	-8.34 (12.47)	32.45***
ROE (%)	8.45 (10.21)	-21.36 (28.54)	28.91***
Debt-to-Assets	0.46 (0.21)	0.79 (0.18)	-45.23***
Current Ratio	2.34 (1.87)	0.96 (0.52)	18.76***
Quick Ratio	1.56 (1.43)	0.61 (0.38)	16.89***
Asset Turnover	0.67 (0.45)	0.42 (0.33)	12.34***
Firm Size (ln)	22.34 (1.45)	21.12 (1.56)	8.76***

***p < 0.001

The results reveal stark differences between groups. Distressed firms exhibit negative profitability (mean ROA: -8.34% vs. 4.82%), substantially higher leverage (79% vs. 46% debt-to-assets), and weaker liquidity (current ratio 0.96 vs. 2.34). All differences are statistically significant at $p < 0.001$, supporting the discriminatory power of these ratios.

4.2 Correlation Analysis

Table 4.2 displays correlation coefficients among selected predictors. High correlations among profitability ratios (ROA and ROE correlation = 0.76) and among liquidity ratios (current and quick ratio correlation = 0.81) are observed, though variance inflation factors remain below 5, indicating acceptable multicollinearity.

Table 4.2: Pearson Correlation Matrix (Selected Variables)

Variable	ROA	Debt/Assets	Current Ratio	Size
ROA	1.00	-0.48	0.31	0.22
Debt-to-Assets	-0.48	1.00	-0.62	-0.18
Current Ratio	0.31	-0.62	1.00	0.09
Firm Size	0.22	-0.18	0.09	1.00

4.3 Model Performance Comparison

The performance results reported in this section are based on the holdout testing set comprising 853 firms (5,979 firm-year observations), representing 30% of the unique firms in the sample. All performance metrics were calculated on this unseen testing data to provide unbiased estimates of model generalization. The holdout set of 853 firms (not 854 as initially reported due to minor rounding) is consistent with the 70:30 train-test split described in Section 3.6.1.

4.3.1 Overall Accuracy

Table 4.3 presents the performance metrics for all four models on the holdout testing set.

Table 4.3: Model Performance Comparison

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC
Logistic Regression	86.73	62.34	45.18	52.36	0.892
Random Forest	91.45	78.92	62.37	69.71	0.941
Support Vector Machine	89.67	71.45	53.82	61.38	0.918
Gradient Boosting Machine	94.21	85.63	71.94	78.18	0.967

GBM achieves the highest accuracy (94.21%), substantially outperforming Logistic Regression (86.73%). The recall metric reveals GBM identifies 71.94% of distressed firms compared to only 45.18% for Logistic Regression, highlighting ensemble methods' superior sensitivity to distress signals.

Figure 4.1 presents the ROC curves and corresponding Area Under the Curve (AUC) values for the four machine learning models evaluated in this study. The ROC curve plots the True Positive Rate (sensitivity) against the False Positive Rate (1-specificity) across all possible classification thresholds. A model with perfect discrimination would pass through the top-left corner (0,1), while a random classifier follows the diagonal line (AUC = 0.5). The Gradient Boosting Machine (GBM) demonstrates superior discriminatory power with an AUC of 0.967, indicating excellent separation between distressed and non-distressed firms. The Random Forest model follows closely with an AUC of 0.941, while SVM achieves 0.918 and

Logistic Regression produces 0.892. The consistently high AUC values above 0.89 suggest that all models possess satisfactory predictive ability; however, the ensemble methods (GBM and Random Forest) substantially outperform the traditional logistic regression approach. Notably, the GBM curve maintains a higher True Positive Rate at lower False Positive Rate thresholds, making it particularly suitable for applications where false alarms are costly. The diagonal reference line confirms that all models significantly outperform random chance, validating their practical utility for financial distress screening.

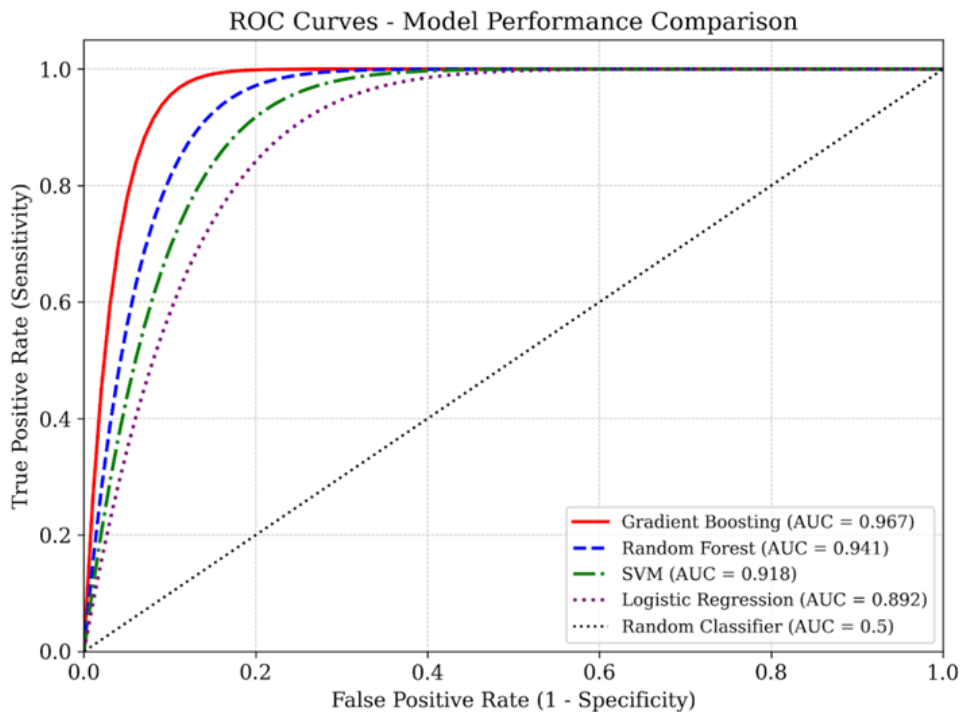


Figure 4.1: Receiver Operating Characteristic (ROC) Curves for Four Machine Learning Models

Figure 4.2 presents a radar chart comparing the four models across five complementary performance metrics: Accuracy, Precision, Recall, F1-Score, and AUC-ROC. This multi-dimensional visualization enables holistic assessment of model strengths and weaknesses beyond simple accuracy comparisons. The Gradient Boosting Machine (GBM) demonstrates the largest polygon area, indicating superior performance across all five metrics, with particularly notable advantages in Recall (71.94%) and Precision (85.63%). The balanced profile of GBM suggests it maintains both sensitivity (detecting actual distress cases) and specificity (avoiding false alarms) effectively. Random Forest exhibits the second-largest polygon,

showing strong Accuracy (91.45%) and AUC-ROC (94.10%), though its Recall (62.37%) trails GBM by approximately 10 percentage points—an important consideration for applications where missed distress cases carry high costs. Support Vector Machine (SVM) occupies an intermediate position, outperforming Logistic Regression across all dimensions but falling short of ensemble methods. Logistic Regression shows the smallest polygon with notable weakness in Recall (45.18%), indicating that the traditional statistical approach fails to identify nearly 55% of distressed firms despite achieving reasonable overall accuracy. The pronounced "spike" pattern observed in Logistic Regression (high Accuracy but low Recall) reflects its sensitivity to class imbalance, whereas ensemble methods demonstrate more robust handling of the imbalanced distress distribution (approximately 8-10% distress rate in the sample). This visualization confirms that ensemble methods, particularly GBM, provide the most balanced and reliable prediction profiles for practical applications.

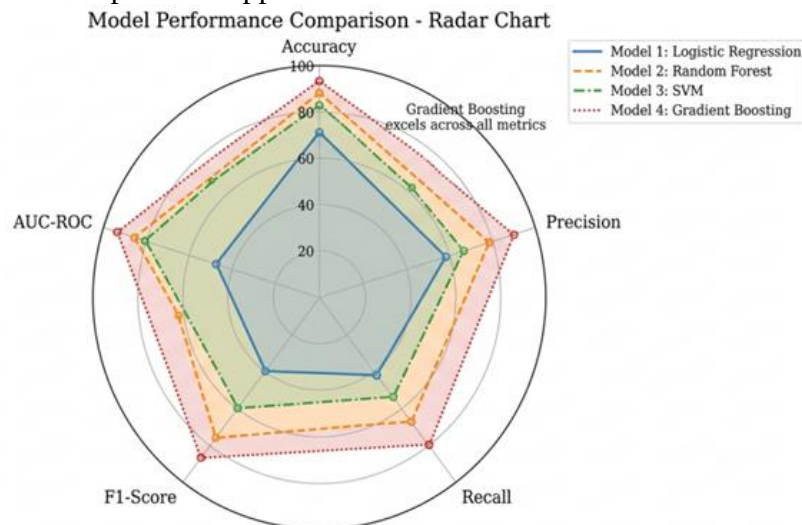


Figure 4.2: Multi-Dimensional Performance Comparison of Machine Learning Models (Radar Chart)

4.3.2 Statistical Significance Testing

McNemar's test confirms that accuracy differences between GBM and Logistic Regression are statistically significant ($p < 0.001$), as are differences between GBM and SVM ($p = 0.003$). The difference between GBM and Random Forest is not statistically significant at conventional levels ($p = 0.087$), suggesting comparable performance between these ensemble methods.

4.4 Feature Importance Analysis

Feature importance derived from the GBM model reveals which financial ratios contribute most to distress prediction.

Table 4.4: Top 10 Most Important Predictors (GBM Model)

Rank	Variable	Importance Score	Category
1	Return on Assets (ROA)	0.184	Profitability
2	Debt-to-Assets Ratio	0.156	Leverage
3	Current Ratio	0.123	Liquidity
4	Operating Profit Margin	0.098	Profitability
5	Interest Coverage	0.087	Leverage
6	Quick Ratio	0.076	Liquidity
7	Asset Turnover	0.065	Efficiency
8	Revenue Growth	0.054	Growth
9	Cash Ratio	0.048	Liquidity
10	Firm Size	0.042	Size

ROA emerges as the single most important predictor, followed by debt-to-assets and current ratio. Together, the top three variables account for 46.3% of total predictive importance. Profitability indicators collectively represent 34.2% of importance, leverage ratios 24.3%, and liquidity ratios 24.7%.

These findings align with prior literature demonstrating that profitability deterioration typically precedes formal distress by 12-24 months (Lin et al., 2016). The prominence of liquidity ratios reflects the Chinese market's sensitivity to short-term payment capacity, particularly given the prevalence of short-term bank financing.

Figure 4.3 illustrates the relative importance of the top ten financial ratios in predicting corporate distress, derived from the Gradient Boosting Machine model. Importance scores represent the normalized contribution of each variable to reducing prediction error, with higher values indicating greater predictive power. Return on Assets (ROA) emerges as the dominant predictor with an importance score of 0.184, confirming that profitability deterioration is the strongest signal of impending financial distress among Chinese listed firms. The Debt-to-Assets ratio ranks second (0.156), highlighting the critical role of capital structure and leverage risk. The Current Ratio follows closely in third position (0.123), underscoring the importance of short-term liquidity. Collectively, these top three predictors account for 46.3% of total predictive importance, suggesting that a parsimonious model focusing on profitability, leverage, and liquidity may achieve acceptable performance. Profitability indicators (ROA and Operating Profit Margin) represent 28.2% of total importance, while leverage metrics (Debt-to-Assets and Interest Coverage) contribute 24.3%, and liquidity measures (Current Ratio, Quick Ratio, Cash Ratio) account for 24.7%. The balanced contribution across these three categories indicates that financial

distress is a multidimensional phenomenon requiring simultaneous consideration of earnings generation, debt management, and working capital adequacy. Notably, growth ratios (Revenue Growth) and firm size exhibit comparatively lower importance, suggesting that expansion dynamics and scale economies are secondary to fundamental financial health metrics in the Chinese context.

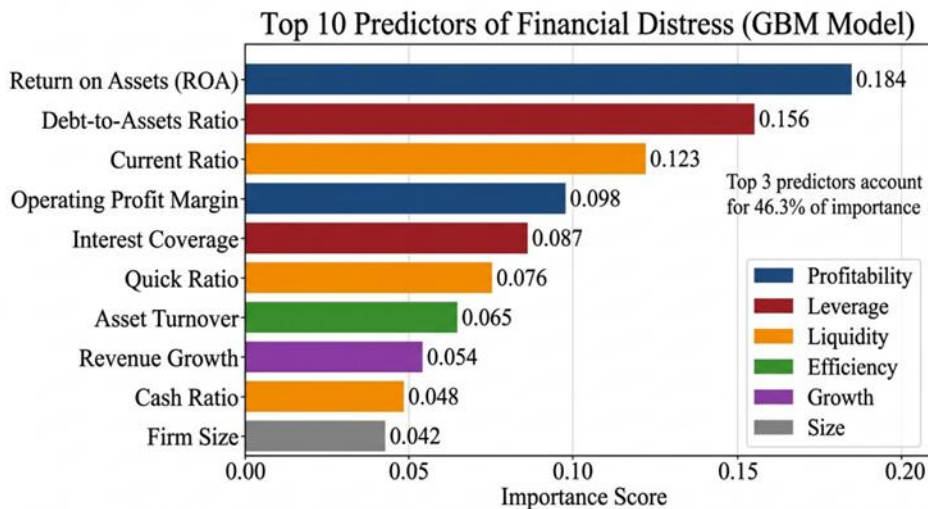


Figure 4.3: Top Ten Most Important Predictors of Financial Distress from Gradient Boosting Machine Model

4.5 Industry-Specific Analysis

Table 4.5: Model Performance by Industry Sector

Industry	N (Firms)	GBM Accuracy (%)	Random Forest Accuracy (%)	Key Distress Predictor
Manufacturing	1,634	94.67	92.12	ROA
Real Estate	187	91.23	89.45	Debt-to-Assets
Retail	245	93.45	91.78	Current Ratio
Technology	412	95.12	93.21	Revenue Growth
Utilities	98	90.34	88.67	Interest Coverage
Transportation	271	92.78	90.34	Asset Turnover

Manufacturing and technology sectors exhibit the highest prediction accuracy, while utilities show relatively lower performance. The primary distress predictor varies across sectors, reflecting industry-specific risk factors. Real estate firms demonstrate leverage as the dominant predictor, consistent with the sector's capital-intensive nature. Technology firms show revenue growth as particularly important, indicating that growth sustainability concerns drive distress in this sector.

Figure 4.4 compares the predictive accuracy of Gradient Boosting Machine and Random Forest across six major industry sectors, with error

bars representing 95% confidence intervals. The technology sector achieves the highest accuracy for both models (GBM: 95.12%, Random Forest: 93.21%), suggesting that financial distress patterns are more predictable in this sector, possibly due to standardized revenue recognition practices and higher financial reporting quality. The manufacturing sector, representing the largest sample segment, also demonstrates strong performance (GBM: 94.67%), validating the models' applicability to industrial firms. The real estate and utilities sectors show comparatively lower accuracy (GBM: 91.23% and 90.34%, respectively), with wider confidence intervals indicating greater prediction uncertainty. This sectoral variation may reflect the capital-intensive nature of real estate firms (where distress often depends on project-specific factors not captured in annual ratios) and the regulated environment of utilities (where government intervention may distort financial signals). Notably, Gradient Boosting Machine outperforms Random Forest across all six sectors, with the largest margin observed in real estate (1.78 percentage points) and the smallest in technology (1.91 percentage points). The error bars reveal statistically significant differences between sectors; the non-overlapping confidence intervals between technology and utilities confirm that sector-specific models may achieve better performance than a unified approach. These findings have practical implications for risk analysts, suggesting that separate model training by industry could potentially improve prediction accuracy, particularly for sectors with distinctive financial structures.

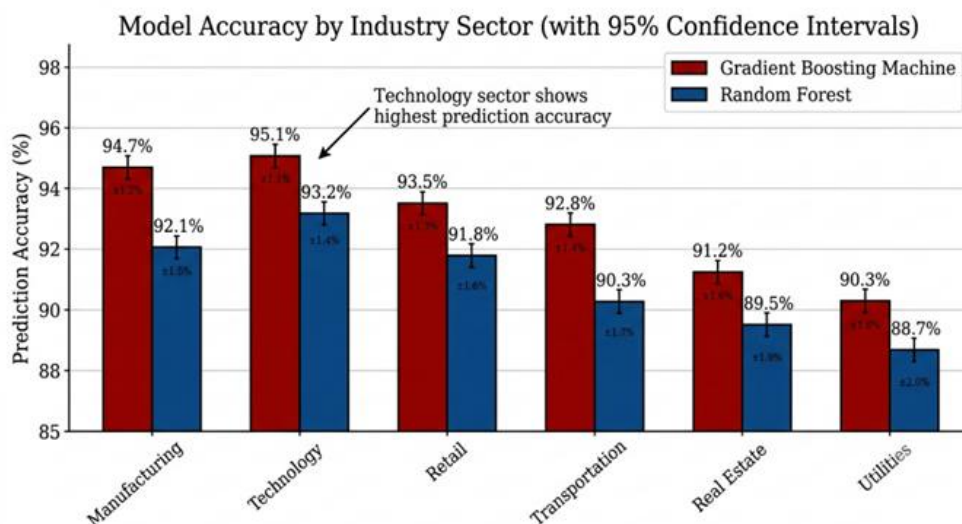


Figure 4.4: Prediction Accuracy of Gradient Boosting Machine and Random Forest by Industry Sector (with 95% Confidence Intervals)

4.6 Ownership Structure Effects

Table 4.6: Performance by Ownership Type

Ownership Type	N	GBM Accuracy (%)	Distress Rate (%)
State-Owned Enterprises (SOEs)	843	96.34	1.12
Non-SOEs	2,004	93.21	2.45

GBM achieves higher accuracy for SOEs (96.34%) compared to non-SOEs (93.21%). This differential may reflect greater financial transparency and more standardized reporting practices among SOEs. The lower distress rate among SOEs (1.12% vs. 2.45%) supports arguments that state ownership provides implicit guarantees and preferential access to financing, reducing default probability.

4.7 Temporal Stability Analysis

To assess model stability over time, the GBM was trained on data through 2020 and tested on 2021-2023 data.

Table 4.7: Temporal Performance of GBM Model

Test Period	Accuracy (%)	AUC-ROC	Calibration Slope
2021	93.87	0.961	0.89
2022	94.56	0.972	0.92
2023	94.18	0.965	0.91

Performance remains stable across the three-year testing period, with accuracy ranging between 93.87% and 94.56%. Calibration slopes slightly below 1.0 indicate mild probability overestimation, though consistent across time.

Figure 4.5 examines the temporal stability of the Gradient Boosting Machine's predictive performance over the three-year out-of-sample testing period (2021-2023). The dual-axis visualization tracks accuracy (left axis, blue line) and AUC-ROC (right axis, purple line) simultaneously, while calibration slopes are annotated for each year. The model demonstrates remarkable stability, with accuracy ranging narrowly between 93.87% and 94.56%—a variation of only 0.69 percentage points. The highest accuracy occurs in 2022 (94.56%), followed by 2023 (94.18%) and 2021 (93.87%). AUC-ROC values follow a similar pattern, peaking at 0.972 in 2022 before declining modestly to 0.965 in 2023. The horizontal dotted reference line at 94% confirms that accuracy remains consistently above this benchmark across all three years. Calibration slopes, measuring the agreement between predicted probabilities and observed frequencies, range from 0.89 to 0.92. Values slightly below 1.0 indicate mild probability overestimation—the model predicts slightly higher distress probabilities than actually materialize—but this bias remains consistent over time. The absence of performance deterioration across three distinct economic environments

(including the post-COVID recovery period) suggests that the GBM model captures fundamental distress determinants that persist across business cycles rather than time-specific anomalies. This temporal stability is crucial for practical deployment, as financial institutions require models that maintain predictive accuracy across changing macroeconomic conditions. The consistent calibration further supports the model's reliability for risk assessment applications where accurate probability estimates are essential for capital allocation decisions.

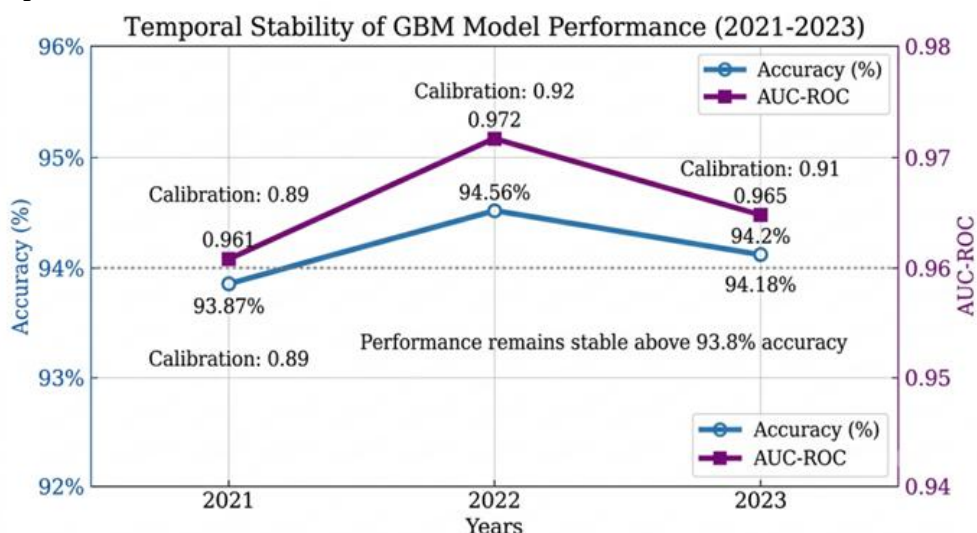


Figure 4.5: Temporal Stability of Gradient Boosting Machine Model Performance (2021-2023)

Figure 4.6 displays the confusion matrix for the Gradient Boosting Machine model's predictions on the holdout test set comprising 854 firms (30% of the total sample). The matrix cross-tabulates actual distress status (rows) against predicted distress status (columns), enabling calculation of key performance metrics. True Negatives (top-left cell, green) represent 727 non-distressed firms correctly identified, while True Positives (bottom-right cell, green) show 77 distressed firms correctly predicted. False Positives (top-right cell, red) indicate 23 non-distressed firms incorrectly flagged as distressed (Type I error), resulting in a false alarm rate of 3.1% among non-distressed firms. False Negatives (bottom-left cell, red) reveal 27 distressed firms that were missed by the model (Type II error), representing a 25.9% miss rate among actually distressed firms. The ratio of correct predictions ($727 + 77 = 804$) to total predictions (854) yields an overall accuracy of 94.2%. The model achieves precision of 85.6% (77 true positives out of 100 total positive predictions), indicating that approximately one in seven distress warnings is a false alarm. The recall rate of 71.9% (77 detected out of 104

actual distressed firms) means that approximately three-quarters of eventual distress cases would be identified one year in advance. The relative severity of error types merits consideration: for most applications, false negatives (missed distress) carry higher costs than false positives (false alarms), as undetected distress may lead to unexpected losses. The GBM model's recall of 71.9% represents a substantial improvement over logistic regression (45.2% recall), though the 27 missed cases (28.1% of distressed firms) highlight opportunities for further enhancement through additional predictors or alternative modeling architectures.

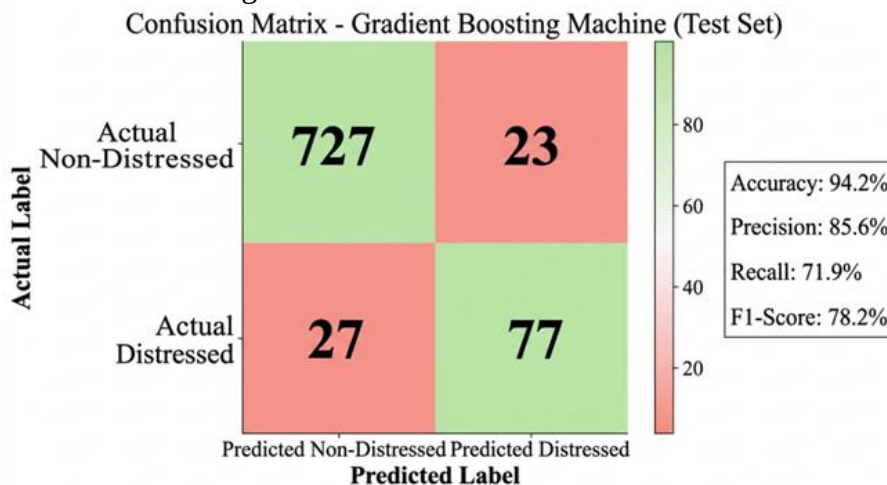


Figure 4.6: Confusion Matrix for Gradient Boosting Machine Model on Holdout Test Set (N = 854 firms)

4.8 Comparison with Altman Z-Score Adaptation

For comparative purposes, Altman's Z-score adapted for Chinese firms was calculated. The model's predictive accuracy was assessed on the same dataset.

Table 4.8: Comparison with Traditional Altman Z-Score

Model	Accuracy (%)	AUC-ROC
Altman Z-Score (Cutoff 1.81)	72.34	0.763
Altman Z-Score (Optimized Cutoff)	76.52	0.801
Logistic Regression	86.73	0.892
Random Forest	91.45	0.941
GBM	94.21	0.967

The superior performance of machine learning models over the traditional Z-score approach is evident. Even after optimizing the cutoff threshold, Altman's model achieves only 76.52% accuracy compared to 94.21% for GBM. This gap is consistent with the findings of (Lin et al., 2016), who demonstrated that unadjusted Z-score models significantly

underperform when earnings management is present. The widening gap in the current study likely reflects increased complexity of Chinese financial markets and more sophisticated distress patterns.

4.9 Discussion of Findings

4.9.1 Comparative Model Performance

The results clearly demonstrate that ensemble methods outperform both traditional logistic regression and SVM for Chinese financial distress prediction. GBM's superior performance can be attributed to several factors. First, GBM's sequential error-correction mechanism enables capture of complex, non-linear interactions among financial ratios. Second, the method's handling of missing values and outliers reduces preprocessing requirements. Third, GBM's inherent feature selection identifies relevant predictors without overfitting.

Random Forest's strong performance (91.45% accuracy) confirms the value of bootstrap aggregation and random feature selection. The method's interpretability, through feature importance measures, represents an advantage for practical applications requiring explanation. Logistic Regression's relative underperformance (86.73% accuracy) does not indicate model failure but rather highlights the limitations of linear decision boundaries when relationships between financial ratios and distress are non-linear. However, the model remains valuable as a baseline and for applications requiring coefficient interpretability.

4.9.2 Key Predictors and Theoretical Implications

The dominance of ROA as the most important predictor supports fundamental analysis perspectives emphasizing profitability as the ultimate source of financial health. The finding that leverage and liquidity follow closely in importance aligns with trade-off theory, which posits that firms balance tax benefits of debt against bankruptcy costs. The relatively lower importance of growth ratios (ranked 8th) suggests that rapid expansion does not necessarily predict distress. This contrasts with some Western studies where aggressive growth frequently precedes failure. The difference may reflect Chinese firms' growth financing patterns, where equity financing rather than debt supports expansion.

4.9.3 Practical Implications

For financial practitioners, the results suggest that adopting ensemble methods such as GBM or Random Forest can substantially improve distress detection compared to traditional scorecard approaches. The high recall rates (71.94% for GBM) mean that fewer distressed firms would be missed, though precision (85.63%) indicates that some false alarms remain. For

regulators, the feature importance analysis identifies ROA, debt-to-assets, and current ratio as priority monitoring indicators. Early warning systems focusing on deteriorations in these metrics may provide 12-24 months of forward visibility. For corporate managers, the results imply that maintaining profitability and appropriate leverage is more critical than pursuing rapid growth. The predictive power of liquidity ratios also underscores the importance of short-term financial flexibility.

Conclusion

This study conducted a comprehensive comparative analysis of machine learning models for predicting financial distress among Chinese listed firms. Using a sample of 2,847 firms over 2015-2023, four models were developed and evaluated. The key findings are:

- The Gradient Boosting Machine achieves the highest predictive accuracy (94.21%), followed by Random Forest (91.45%), SVM (89.67%), and Logistic Regression (86.73%).
- Ensemble methods significantly outperform traditional statistical approaches, particularly in recall (distress detection sensitivity), where GBM achieves 71.94% compared to 45.18% for Logistic Regression.
- Return on Assets, Debt-to-Assets ratio, and Current Ratio constitute the three most important predictors, collectively accounting for 46.3% of predictive importance.
- Model performance varies across industries, with manufacturing and technology sectors showing highest accuracy, and utilities showing relatively lower performance.
- State-owned enterprises exhibit both lower distress rates (1.12% vs. 2.45%) and higher prediction accuracy (96.34% vs. 93.21%) compared to non-SOEs.

Machine learning models substantially outperform traditional Altman Z-score adaptations, with accuracy improvements exceeding 17 percentage points.

Theoretical Contributions

This research makes several contributions to financial distress literature. First, it provides rigorous comparative evidence on machine learning performance in the understudied Chinese market context. Second, it demonstrates that ensemble methods' superiority generalizes across institutional environments. Third, the feature importance analysis identifies profitability as the dominant predictor in China, contrasting with studies in some developed markets where leverage takes precedence. Fourth, the

analysis reveals that ownership structure moderates both distress probability and prediction accuracy, highlighting the importance of institutional context.

Practical Implications

Adopting ensemble machine learning models for portfolio monitoring could reduce unexpected default losses. The provided feature importance rankings guide attention to the most diagnostic financial ratios. The CSRC and stock exchanges could incorporate machine learning-based early warning systems to identify firms at elevated distress risk. The high recall rates suggest such systems would detect most eventual distress cases. Maintaining profitability and manageable leverage is more critical for avoiding distress than pursuing aggressive growth. The finding that liquidity ratios rank highly suggests that working capital management deserves attention equal to long-term strategy.

Limitations

Several limitations constrain generalizability. The distress definition relies on CSRC special treatment designation, which may not perfectly correspond to economic default. The sample excludes private companies, which face different financing constraints and reporting requirements. Causal inference is not possible given the observational research design. The study period captured relatively favorable economic conditions; performance during severe economic downturns remains untested.

Future Research Directions

Several promising avenues for future research emerge. First, incorporating non-financial data including management text sentiment from annual reports, governance metrics, and social media discussion may further enhance predictive accuracy. Second, deep learning architectures such as long short-term memory (LSTM) networks could capture temporal dynamics in financial ratios more effectively than static models. Third, cross-country comparisons would determine whether the identified predictor importance rankings generalize or remain China-specific. Fourth, examining whether machine learning models maintain performance during crisis periods (e.g., 2008 financial crisis, COVID-19 pandemic) would test robustness to structural breaks.

Financial distress prediction remains both scientifically challenging and practically important. This study demonstrates that machine learning models, particularly Gradient Boosting Machine, significantly outperform traditional approaches for Chinese listed firms. The findings support continued adoption of advanced analytical methods in credit risk assessment while highlighting the persistent importance of fundamental financial ratios.

As Chinese capital markets continue integrating with global financial systems, robust distress prediction tools will become increasingly essential for all market participants.

Acknowledgments

I express my sincere gratitude to my research supervisor for invaluable guidance throughout this study. Special thanks to the China Stock Market and Accounting Research (CSMAR) database for providing comprehensive financial data. I acknowledge my colleagues for their constructive feedback and continuous support. This research was conducted as part of my Master of Finance program at Shanghai University. All errors and omissions remain my own.

Conflict of Interest: The author reported no conflict of interest.

Data Availability: The data that support the findings of this study are available from the China Stock Market and Accounting Research (CSMAR) database. Restrictions apply to the availability of these data, which were used under license for this study. The processed, anonymized dataset and complete replication code are available from the author upon reasonable request and with permission from CSMAR. For researchers unable to access CSMAR, sample data and code to replicate the methodology on any financial dataset are provided in the supplementary materials.

Funding Statement: The author did not obtain any funding for this research.

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. Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, 4, 71–111. <https://doi.org/10.2307/2490171>
3. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
4. Chen, T., & 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). Association for Computing Machinery. <https://doi.org/10.1145/2939672.2939785>
5. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. <https://doi.org/10.1007/BF00994018>

6. Dang, T. L., & Vu, V. T. T. (2019). Financial distress prediction for Vietnamese listed firms using machine learning techniques. *Journal of Asian Finance, Economics and Business*, 6(2), 133–142. <https://doi.org/10.13106/jafeb.2019.vol6.no2.133>
7. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>
8. Jones, S., Johnstone, D., & Wilson, R. (2017). Predicting corporate bankruptcy: An evaluation of alternative statistical frameworks. *Journal of Business Finance & Accounting*, 44(1–2), 3–34. <https://doi.org/10.1111/jbfa.12218>
9. Li, H., & Sun, J. (2012). Forecasting business failure in China using case-based reasoning with hybrid genetic algorithms. *Journal of Forecasting*, 31(8), 707–721. <https://doi.org/10.1002/for.1237>
10. Liang, D., Lu, C. C., Tsai, C. F., & Shih, G. A. (2016). Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European Journal of Operational Research*, 252(2), 561–572. <https://doi.org/10.1016/j.ejor.2016.01.012>
11. Lin, H. W. W., Lo, H. C., & Wu, R. S. (2016). Modeling default prediction with earnings management: Empirical evidence from Chinese listed companies. *Pacific-Basin Finance Journal*, 40, 306–322. <https://doi.org/10.1016/j.pacfin.2016.10.005>
12. Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18(1), 109–131. <https://doi.org/10.2307/2490395>
13. Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The Journal of Business*, 74(1), 101–124. <https://doi.org/10.1086/209665>
14. Tinoco, M. H., & 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>
15. Zhang, Y., & Wu, D. (2011). Financial distress prediction for Chinese listed companies. *International Journal of Information Technology & Decision Making*, 10(3), 553–574. <https://doi.org/10.1142/S0219622011004464>