

Can Capital Market Information Predict Corporate Risk Before Financial Deterioration? Evidence from Sci-Tech Firms

SHEN Chuan-he

Shandong Women's University

SHEN Xu

Shandong Women's University

CHEN Xu-feng

19859916956@163.com

Shandong University of Science and Technology

Article

Keywords: science and technology innovation enterprises, Credit Risk, conditional volatility, information disclosure, machine learning, nonlinear relationship

Posted Date: October 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11178131/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

There is still a lack of sufficient empirical evidence on whether capital market information can identify credit risk in sci-tech firms before the deterioration of conventional financial indicators. Using quarterly data from 454 STAR Market-listed firms over 2019–2024, we predict each firm's credit-risk status. We then develop a multi-source risk early-warning system that integrates financial indicators, market behavior, GARCH conditional volatility, network linkages, innovation characteristics and information disclosure behavior. Time-series out-of-sample validation, LightGBM, and the SHAP method are employed to examine the incremental predictive power of this system. After GARCH conditional volatility is added, the model's AUC increased from 0.7905 to 0.8304, indicating that conditional volatility provides incremental information beyond historical volatility, and the relationship between conditional volatility and credit risk is nonlinear and state-dependent. Network linkages does not provide additional predictive gains on its own, while information disclosure behavior yields the largest AUC gain of 0.093. SHAP results further show that R&D investment, receivables turnover, cash flow quality, and consecutive gaps in R&D information are important risk signals. LightGBM achieved an AUC of 0.875 on the test set. The findings suggest that credit risk in sci-tech firms exhibits dynamic and nonlinear patterns and is driven by information from multiple sources.

1. Introduction

Sci-tech firms have the characteristics of high R&D investment, asset-light operations, and uncertainty about future cash flows. Market expectations about their credit risk may change before conventional financial indicators deteriorate. Conventional credit risk research has relied mainly on financial ratios, distance-to-default, and Logit models [1–3], but low-frequency financial information is often insufficient to capture changes in the risk profile of sci-tech enterprises in a timely manner. As the information efficiency of capital markets has improved, stock prices, return volatility, and trading behavior have become important sources of information for identifying corporate risk [4–6]. In particular, GARCH conditional volatility can characterize time-varying risk and volatility clustering, and is more likely than historical volatility to reflect the market's dynamic expectations of future uncertainty [7–8].

At the same time, the credit risk of technology-based innovative enterprises is not determined solely by financial conditions. R&D investment, innovative personnel, and patent accumulation influence firms' future value creation, but the high investment, long cycle, and high uncertainty associated with R&D activities may also generate financing pressure [9–11]. More importantly, R&D information has a high degree of disclosure heterogeneity, and the consecutive gaps in R&D information itself may include information opacity and governance risks. Therefore, it is difficult to fully capture the credit risk in sci-tech firms by relying solely on financial indicators or individual market-based measures.

Several gaps remain in the existing literature. First, relatively little attention has been paid to the forward-looking incremental value of capital market information from the perspective of dynamic conditional volatility. Second, capital market information, innovation characteristics, and information disclosure behavior are usually examined separately, and there is a lack of a unified multidimensional framework for

credit-risk identification. Third, it is difficult for traditional linear models to capture nonlinear and state-dependent relationships among risk factors, and machine-learning studies rarely provide further insight into the underlying economic mechanisms.

Based on this, the present study takes companies listed on the Sci-Tech Innovation Board (STAR Market) from 2019 to 2024 as the research sample and constructs a credit risk early-warning system that integrates conditional volatility, investor behavior, network connectivity, innovation characteristics, and information disclosure behavior. Using time-series out-of-sample validation and the LightGBM algorithm, we assess the incremental predictive value of different information dimensions. On this basis, SHAP and volatility-stratification methods are further employed to identify the nonlinear risk structure and key driving factors.

This paper makes two main contributions. First, it examines the incremental value of conditional volatility relative to conventional financial indicators and historical volatility from the perspective of risk discovery. Second, it incorporates innovation characteristics and information disclosure behavior into an interpretable machine-learning framework to reveal the nonlinear and multidimensional information-driven mechanisms of credit risk in sci-tech firms.

2. Theoretical Mechanism and Research Design

2.1. Theoretical Mechanisms and Research Hypotheses

The capital market play a continuous role in information discovery and risk pricing. When investors expect a firm's future cash flows to deteriorate or business uncertainty to rise, risk repricing may first manifest as return volatility and abnormal trading activities. Therefore, GARCH conditional volatility, RSJ and abnormal turnover rate can provide forward-looking risk signals beyond conventional financial information from the market level. Capital market risk expectations may also affect firms' credit conditions through financing costs and the availability of funds. This leads to the following hypothesis:

H1: Capital market dynamic information can improve the out-of-sample predictive power for the credit risk of sci-tech innovation enterprises.

The value of sci-tech firms depends heavily on their R&D activities and innovation outcomes. Continuous R&D investment, the accumulation of R&D personnel and patent output can strengthen a firm's long-term competitive position, R&D activities also have the characteristics of high costs, long payback periods and limited collateral value. Their implications for credit risk depend on the cash-flow capacity of the enterprise and the efficiency of converting innovation into commercial outcomes. In addition, innovation-related information disclosure may influence external investors' assessment of a firm's future value and risk. Both innovation capability and its disclosure status may thus serve as important sources of information for credit risk. This leads to the following hypothesis:

H2: Innovation characteristics and information disclosure behavior provide significant incremental value for credit-risk identification.

Capital market information and innovation characteristics may interact rather than operate independently. The same level of market volatility may have different economic implications for firms with stronger innovation capabilities and those with weaker innovation capabilities, and the risk implications of innovation investment also depends on financing conditions and market expectations. A nonlinear and state-dependent relationship may therefore exist between the two. This leads to the following hypothesis:

H3: There is a nonlinear and state-dependent relationship between capital market information and innovation characteristics, their interaction enhances the credit-risk identification of sci-tech firms.

2.2. Samples, Variables, and Prediction Strategies

This study examines firms listed on the Science and Technology Innovation Board (STAR Market) from July 2019 to December 2024, using transaction data from CSMAR and financial data from Wind. After screening, we obtain 8691 firm-quarter observations from 454 firms. According to the forward-looking prediction framework, all explanatory variables are lagged by one quarter, the credit-risk status in the following quarter is the prediction target, and the observations of future risk labels and lag variables that could not be constructed are eliminated. This leaves 7,777 valid observations, including 393 risk observations, or 5.05% of the sample. This design avoids future information leakage caused by random sample splitting.

Rather than equating these events with conventional debt default, we define the outcome as a composite corporate risk state that captures financial Distress risk, regulatory inquiries, and non-standard audit opinions. Core market variables include GARCH conditional volatility, RSJ and abnormal turnover rate, and innovation characteristics include R&D investment, R&D personnel and patent indicators, and we also construct measures of continuity in R&D information disclosure. The models control the current ratio, leverage, operating profit margin, receivables turnover rate, operating revenue growth rate and cash flow quality. Missing innovation-related information is not simply mean-imputed, but its potential information content is retained through the information gaps indicators.

The model adopts LightGBM and divides the training, validation and test sets in chronological order. Class weighting is only employed in the training stage to preserve the actual risk proportion in the test set. The model performance is mainly evaluated by indicators such as AUC, KS, F1, PR-AUC and Brier, and Random Forest, XGBoost and Elastic Net are employed as benchmarks.

3. Empirical Results

3.1. The incremental predictive value of conditional volatility

Table 1 provides an overview of the final sample. The average leverage ratio is 28.70%, whereas the mean operating profit margin is -0.424. R&D investment, R&D personnel and patent-related variables have a high rate of missing values. Together, these features suggest substantial differences across sci-tech firms in both profitability and the availability of innovation-related information. GARCH modeling results further show that, among the 454 firms, 64.55% follow the standard GARCH specification, 32.60% follow EGARCH, and 2.84% follow GJR-GARCH. The predominance of Student's t and skewed t distributions suggests that the stock returns of sci-tech innovation enterprises exhibit pronounced fat tails and, in certain cases, asymmetric features.

Table 1
Descriptive Statistics

Variable	Sample Size	Mean	Standard Deviation	Minimum	Median	Maximum	Missing Rate
Panel A: Dependent Variable							
Risk State ($t + 1$)	8176	0.0472	0.2121	0	0	1	0%
Panel B: Volatility Measures							
Historical Volatility (22-Day)	5675	0.5548	0.2435	0.2012	0.5044	1.513	30.59%
Historical Volatility (63-Day)	5603	0.5906	0.2728	0.2216	0.5287	1.7729	31.47%
GARCH Conditional Volatility	7722	0.4322	0.267	0.0248	0.4911	1.0591	5.55%
Panel C: Financial Indicators							
Debt-to-Asset Ratio	7722	0.287	0.1866	0.0293	0.2482	0.7904	5.55%
Current Ratio	7722	5.9391	6.5881	0.784	3.4177	37.2575	5.55%
Operating Profit Margin	7722	-0.4236	3.7592	-33.4398	0.1278	0.6042	5.55%
Accounts Receivable Turnover Ratio	7722	5.255	10.2234	0.1139	2.4701	78.0695	5.55%
Net Profit Cash Coverage Ratio	7722	-1.4334	9.1548	-71.1096	0.4594	12.1996	5.55%
Operating Revenue Growth Rate	7722	0.1769	0.637	-0.8204	0.0644	3.6898	5.55%
Panel D: Sci-Tech Innovation Characteristics							
<p>Note: The sample size for the dependent variable (risk state at $t + 1$) is 8,176. The effective sample size varies across explanatory variables due to missing observations, with 7,722 observations available for most financial and network variables. All continuous variables are winsorized at the 1st and 99th percentiles. Both historical volatility and GARCH conditional volatility are annualized. PageRank indicators are standardized (Z-scores) to measure network centrality. Due to limitations in information disclosure, sci-tech innovation indicators exhibit relatively high missing rates.</p>							

Variable	Sample Size	Mean	Standard Deviation	Minimum	Median	Maximum	Missing Rate
Panel A: Dependent Variable							
R&D investment ratio	2345	32.3727	130.1709	2.6795	10.14	1139.2685	71.32%
Proportion of R&D personnel	2019	30.0973	18.1944	7.298	24.25	85.669	75.31%
Number of patents	1368	50.2474	82.2612	1	25	541.96	83.27%
Panel E: Behavioral Factors							
RSJ skewness	5591	0.0227	0.2102	-0.4741	0.0241	0.5115	31.62%
Abnormal turnover rate	5646	-0.0052	0.1447	-0.4723	0.0024	0.3492	30.94%
Panel F: Network Centrality							
Composite network influence	7722	-0.2743	0.8274	-3	-0.1531	1.0961	5.55%
Industry PageRank	7722	-0.2724	1.4004	-7.2922	0	1.6926	5.55%
Market capitalization PageRank	7722	-0.3408	1.0739	-3.2932	0.0062	0.9357	5.55%
Volatility PageRank	7722	-0.3506	1.4259	-7.2922	0	1.4754	5.55%
<p>Note: The sample size for the dependent variable (risk state at $(t + 1)$) is 8,176. The effective sample size varies across explanatory variables due to missing observations, with 7,722 observations available for most financial and network variables. All continuous variables are winsorized at the 1st and 99th percentiles. Both historical volatility and GARCH conditional volatility are annualized. PageRank indicators are standardized (Z-scores) to measure network centrality. Due to limitations in information disclosure, sci-tech innovation indicators exhibit relatively high missing rates.</p>							

More importantly, the mean lagged conditional volatility of firms that subsequently experience a risk event is 0.486, compared with 0.429 for non-risk firms, a difference of 0.057 ($t = 5.02$). This suggests that firms subsequently experiencing a risk event exhibit higher market uncertainty before the event occurs.

To test whether conditional volatility has incremental predictive value beyond historical volatility. To this end, we estimate three nested models, with the results reported in Table 2.

Table 2
Comparison of Predictive Performance Across Nested Models

Model	Feature Set	AUC	Δ AUC	DeLong.p.value
M1	Financial indicators	0.7908	—	—
M2	M1 + historical volatility	0.7660	-0.0248	< 0.001***
M3	M2 + GARCH conditional volatility	0.7854	0.0194	< 0.001***

Note: ***p < 0.001. Δ AUC denotes the AUC of the current model minus that of the preceding model. M1 is the Altman (1968)-style financial benchmark. M2 adds 22-day and 63-day historical volatility. M3 further incorporates the GARCH conditional volatility. In order to ensure that comparison across nested feature sets is not confounded by model complexity, this table reports the LightGBM results using the same strict regularization settings.

Table 2 further provides out-of-sample incremental evidence. The financial benchmark (M1) yields an AUC of 0.7908. Adding 22-day and 63-day historical volatility reduces the AUC to 0.7660 (Δ AUC = -0.0248, $p < 0.001$). With GARCH conditional volatility added in M3, the AUC rises to 0.7854 (Δ AUC = 0.0194, $p < 0.001$). These results suggest that GARCH conditional volatility provides incremental predictive information beyond historical volatility, although its contribution does not fully restore the performance of the financial benchmark.

3.2. Nonlinear and Threshold Evidence

To evaluate whether nonlinear interaction effects exist among the risk factors, this study first compares the predictive performance of the linear model with that of the nonlinear model; detailed results are presented in Table 3.

Table 3
Comparison of Predictive Performance between Linear and Nonlinear Models

Model	Model type	AUC	AUC 95%CI	Brier	DeLong p	NA handling
Elastic Net	Linear	0.723	[0.678, 0.768]	0.286	—	Imputation (training set median)
LightGBM	Nonlinear	0.785	[0.744, 0.827]	0.079	< 0.001***	Retained NA

Note: ***indicates $p < 0.001$. Both models use the same base feature set, including financial indicators, historical volatility, and GARCH conditional volatility. Elastic Net imputes missing values using statistics from the training set, whereas LightGBM retains the original missing values. The AUC confidence intervals are reported at the 95% confidence level.

The results in Table 3 show that the two models perform quite differently despite being estimated with the same feature set. Elastic Net achieves an AUC of 0.723, while LightGBM reaches 0.785. The DeLong test confirms that this difference is statistically significant at the 0.1% level ($p < 0.001$). In addition,

LightGBM achieves a substantially lower Brier score, indicating better predictive accuracy in terms of probability estimation. These results suggest that the credit risk of sci-tech firms is not adequately captured by a simple linear combination of risk factors. Nonlinear relationships and higher-order interactions contain important predictive information.

To illustrate the interaction between volatility and financial indicators, this paper computes SHAP values [12] for each feature in the test set. We further use SHAP dependence plot to investigate the marginal contribution of GARCH conditional volatility to the model's predictions. Based on this analysis, we plot the SHAP dependence relationship between conditional volatility and the debt-to-asset ratio.

As shown in Fig. 1, the SHAP contribution of conditional volatility exhibits a nonlinear response pattern rather than a monotonically increasing relationship with volatility levels. At relatively low volatility levels, its marginal contribution remains limited, indicating a weak impact on the model output. As volatility rises into the intermediate range, the SHAP value increases rapidly, suggesting that changes in conditional volatility become an important driver of risk perception. However, after reaching a certain threshold, the marginal effect gradually stabilizes, with the contribution curve approaching a plateau at higher volatility levels. This indicates that conditional volatility primarily functions as a risk-state indicator, whose influence intensifies during volatility transitions but does not continue to amplify risk linearly under extreme conditions.

To further test whether there is a simple linear or monotonic relationship between conditional volatility and credit risk, we use the test set, which is kept completely separate from model training, and divide the observations into five volatility quintiles (Q1 to Q5, from lowest to highest) according to the within-quarter cross-sectional rank of conditional volatility. The within-quarter ranking is adopted because the absolute level of GARCH conditional volatility exhibits a pronounced time drift over the sample period, and directly applying fixed full-sample or training-set thresholds would generate severe distributional bias in the test set. Table 4 reports the results.

Table 4
Credit Risk Incidence by Conditional Volatility Tier

Volatility quintile (within-quarter)	N	Number of risk events	Risk event rate (%)	OR vs Q1	p-value vs Q1
Q1 (lowest 20%)	450	15	3.33	—	—
Q2	450	17	3.78	1.14	0.718
Q3	455	29	6.37	1.97	0.044**
Q4	450	32	7.11	2.23	0.016**
Q5 (highest 20%)	455	21	4.62	1.41	0.396
Note: The volatility quintile is constructed based on the within-quarter cross-sectional rank (0–1) of conditional volatility, which removes the time drift in the absolute level of GARCH conditional volatility. The risk event rate is defined as the number of risk events in $t + 1$ in the group to the number of valid observations. Odds ratios (OR) and p-values come from Fisher's exact tests with Q1 as the reference group. ** $p < 0.05$. The quintile grouping is only employed for out-of-sample stratified analysis on the test set, and is not employed for model training or parameter selection.					

Table 4 provides further stratified evidence. Risk event rates rise from 3.33% in Q1 to 7.11% in Q4, with Q4 significantly exceeding Q1 (OR = 2.23, $p = 0.016$) and Q3 also significantly higher than Q1 (OR = 1.97, $p = 0.044$). In contrast, Q5 declines to 4.62% and does not differ significantly from either Q4 ($p = 0.12$) or Q1 ($p = 0.396$). A quadratic Logit specification yields a negative squared-term coefficient, consistent in sign with an inverted-U pattern, but it is not statistically significant ($p = 0.167$), possibly due to the limited number of risk events in the test set ($n = 114$). We therefore do not claim a strict inverted-U relationship, but characterize the result as evidence of interval heterogeneity in the risk implications of conditional volatility, with the highest risk incidence occurring in the medium-high volatility interval.

3.3. The Incremental Value of Networks, Innovation, and Information Disclosure

We further construct a cross-firm volatility network using conditional volatility, market capitalization, and industry information. Table 5 shows the incremental predictive value of network centrality.

Table 5
The Incremental Predictive Value of Network Centrality

Model	Feature Set	AUC	Δ AUC	DeLong p-value
Model A	All Non-Network Features	0.8901	—	—
Model B	Model A + Network Centrality	0.8769	-0.0132	< 0.001***
Note: Δ AUC = Model B AUC – Model A AUC; *** $p < 0.001$.				

The results in Table 5 show a decline in AUC from 0.8901 to 0.8769 after network centrality is added, corresponding to a Δ AUC of -0.0132. The result shows that the network centrality does not provide an

independent out-of-sample discrimination gain. A more reasonable explanation is that some of its information has been absorbed by the firm's own conditional volatility. Therefore, network linkages is of economic significance, but the marginal prediction value of its topological measures is limited.

To better assess the incremental contribution of GARCH conditional volatility and its derived features, we conduct stratified tests by introducing sci-tech-specific indicators and behavioral factors in different feature layers. The results are presented in Table 6.

Table 6
GARCH Hierarchical Regression Results

Layer	Feature Dimension	Number of New Features	Cumulative Number of Features	AUC	Theoretical Basis
Layer 1	Financial Indicators	6	6	0.791	Five-factor framework [1]
Layer 2	Market Behavior Factors	2	8	0.768	Behavioral finance [13]
Layer 3	GARCH Volatility	1	9	0.789	GARCH volatility [7]
Layer 4	Network Centrality	6	15	0.781	Network theory [14]
Layer 5	Sci-Tech Innovation Indicators	3	18	0.792	Signaling theory [15]
Layer 6	Information Disclosure Behavior	8	26	0.886	Voluntary disclosure theory [16]
Note: Δ AUC is defined as the AUC of the current layer minus that of the preceding layer. Cumulative increase = 0.095. Altman (1968) represents the conventional Z-score framework.					

Table 6 further reveals the incremental contributions of different information dimensions. Financial indicators provide a baseline predictive performance (AUC = 0.791). Adding market behavior factors reduces the AUC to 0.768, while incorporating GARCH conditional volatility increases it to 0.789. The inclusion of network centrality further reduces the AUC to 0.781, whereas sci-tech innovation indicators yield a limited improvement, raising the AUC to 0.792. In contrast, information disclosure behavior substantially improves predictive performance, increasing the AUC by 0.094 to 0.886 and emerging as the largest source of incremental information among all feature dimensions.

3.4. Model Interpretation and Risk Drivers

Figure 2 shows the mean absolute SHAP values calculated from the test set, and the SHAP results further reveal the economic interpretation of risk identification.

As shown in Fig. 2, the mean absolute SHAP value of maximum R&D expenditure is the highest (0.3650), followed by accounts receivable turnover and net profit cash content, while consecutive missing R&D disclosure quarters ranks fourth (0.1335). GARCH conditional volatility and 22-day historical volatility have importance values of 0.0301 and 0.0284, respectively. These results suggest that market volatility contributes additional predictive information, but R&D investment, operating efficiency, cash flow quality, and information disclosure remain more important sources of information for credit-risk identification.

Feature importance rankings, however, do not indicate the direction of a feature's effect on predicted risk. Figure 3 provides this information through the SHAP beeswarm plot. It shows how the values of the top 10 features are associated with changes in the predicted probability of a risk event. The horizontal axis represents the SHAP value, with zero indicating no contribution. Extending to the right means increasing the probability of a risk event, and extending to the left means reducing the probability of a risk event. The color of each point represents the feature value: yellow indicates a higher value and purple a lower value.

The SHAP summary plot provides further evidence on the direction and heterogeneity of these effects. Higher accounts receivable turnover and operating profit margin are generally associated with negative SHAP values, suggesting that stronger operational efficiency and profitability help reduce credit risk. In contrast, higher net profit cash content ratio also tends to contribute to lower risk exposure, highlighting the risk-mitigating role of cash flow quality. Consecutive gaps in R&D disclosure and patent disclosure are concentrated in the positive SHAP region, indicating that information discontinuity and reduced transparency serve as important risk signals. The SHAP distribution of maximum R&D expenditure-to-revenue ratio exhibits substantial dispersion across both positive and negative regions, suggesting that innovation investment has differentiated effects on credit risk depending on its scale and context. Meanwhile, current ratio and patent-related indicators show relatively moderate but differentiated contributions. Overall, the credit risk of sci-tech firms presents the driving structure of “operational capability and cash flow resilience as risk-mitigating forces-innovation investment heterogeneity-information opacity as a risk signal-market and financial volatility as complementary sources of risk information”.

4. Predictive Performance and Economic Value

4.1. Out-of-Sample Performance

After establishing the need for nonlinear modeling, we next examine the final out-of-sample performance of the multidimensional early-warning system developed in this study. We compare four widely employed machine learning algorithms to identify the model that offers the most suitable overall performance. Table 7 reports their full set of test-set metrics.

Table 7
Comparison of Out-of-Sample Predictive Performance Across Models

Model	AUC	95% CI	KS	Precision	Threshold	Recall	F1	Brier
LightGBM	0.8833	[0.8562, 0.9105]	0.6433	0.1661	0.273	0.8772	0.2793	0.0665
XGBoost	0.8702	[0.8382, 0.9022]	0.5703	0.221	0.504	0.7018	0.3361	0.2505
Random Forest	0.9059	[0.8804, 0.9314]	0.6702	0.2065	0.111	0.8421	0.3316	0.0419
Elastic Net	0.7571	[0.7182, 0.7961]	0.4014	0.093	0.667	0.8333	0.1673	0.3999
<p>Note: AUC is the area under the ROC curve for the test set. The AUC and its 95% confidence interval are estimated using the DeLong method. The lower the Brier score, the better the probability calibration. All models use class-weighted loss to address class imbalance. LightGBM and XGBoost retain missing values, allowing the algorithms to exploit potentially informative missing-data patterns. Random Forest and Elastic Net use median imputation based on the training sample.</p>								

As shown in Table 7, Random Forest achieves the highest AUC and KS statistic, followed by LightGBM, with an AUC of 0.883 and a KS statistic of 0.643. Both models outperform XGBoost and Elastic Net in terms of overall discrimination. LightGBM also achieves the highest recall among the four models, while its Brier score indicates substantially better probability calibration than XGBoost and Elastic Net. Although Random Forest performs best on several conventional predictive metrics, the comparison also involves an important difference in missing-data treatment. Random Forest and Elastic Net use median imputation, whereas LightGBM and XGBoost retain missing values. Given this study's premise that non-disclosure may itself convey information about firms' underlying risk, retaining missingness allows LightGBM to exploit potentially informative disclosure patterns rather than replacing them with imputed values. Accordingly, LightGBM is selected as the main model based on the joint consideration of discriminatory performance, risk-event detection, probability calibration, and consistency with the study's treatment of disclosure-related information, rather than on the claim that it uniformly outperforms all competing models.

4.2. Economic Value

Statistical metrics alone are insufficient to fully assess the practical value of a machine-learning model in real-world investment and financing decisions. To assess the economic significance of our LightGBM model, we divide the sample firms into five risk groups (Q1 to Q5), from low to high risk. We then construct risk-sorted portfolios based on these groups. Table 8 reports the realized risk-event rate and risk-capture efficiency for each group.

Table 8
Portfolio Risk Stratification Analysis

Risk Group	Number of Firms	Mean Predicted Probability	Actual Default Rate(%)	Risk Ratio	Cumulative Risk Share(%)
Q1	452	0.0806	0.00	0.00	0.00
Q2	452	0.1293	0.66	0.13	2.63
Q3	452	0.1919	1.10	0.22	7.02
Q4	452	0.2593	3.54	0.70	21.05
Q5	452	0.3435	19.91	3.95	100.00
Total/Average	2260	0.2010	5.04	1.00	100.00
<p>Note: The risk ratio is defined as the ratio of the realized risk-event rate in each group to the overall risk-event rate in the sample, which is 5.04%. The cumulative risk share indicates the cumulative proportion of realized risk events captured by the groups up to each risk level. The Kendall's tau-b coefficient between portfolio rankings and realized risk-event rates is 1.000 (exact two-sided ($p = 0.0167$)), indicating a strictly increasing relationship across the five portfolios. Percentages and ratios may differ slightly due to rounding.</p>					

Table 8 presents the realized risk-event rates across the five portfolios formed by sorting firms according to their predicted risk probabilities. The realized risk-event rate increases monotonically from 0.00% in Q1 to 19.91% in Q5, compared with an overall sample rate of 5.04%. The highest-risk portfolio therefore exhibits a risk-event rate approximately 3.95 times the sample average, whereas no risk events are observed in the lowest-risk portfolio. The Kendall's tau-b coefficient between portfolio rankings and realized risk-event rates is 1.000, with an exact two-sided p-value of 0.0167. These results indicate that the model generates a clear and economically meaningful ordering of realized risk across portfolios, complementing its overall discriminatory performance with informative risk stratification.

The classification errors in financial risk control have highly asymmetrical economic consequences. A False negative may allow a genuine credit-risk event to go undetected, potentially resulting in substantial credit losses, whereas a false positive mainly increases the cost of additional due diligence. Therefore, the key objective of a risk-control model is to maintain false positives at a manageable level while keeping false negatives very low.

As shown in Fig. 4, at the Youden threshold, the model produces only 14 false negatives, accounting for 0.62% of the test set. This indicates that, in a highly imbalanced credit-risk setting, the model can achieve higher risk-capture capability while accepting some false positives.

4.3. Robustness Checks

Rolling time windows, alternative forecast horizons, and instrumental-variable tests generally support the main conclusions. The rolling-window results shows that the model maintains a high AUC across the regular estimation windows. After the prediction period is extended to $t + 2$ and $t + 3$, the model still

maintains strong discriminatory power, suggesting that the predictive signal is not merely driven by short-term market noise.

In the instrumental-variable analysis, Industry-level average volatility and the VIX are strongly associated with the endogenous volatility measure. The weak-instrument tests support instrument relevance, and the direction of the interaction term in the second stage is consistent with the benchmark results. The above results show that the incremental predictive value of conditional volatility and its nonlinear relationship with financial vulnerability do not depend on a single time window or model setting.

5. Conclusions and Policy Implications

Based on the quarterly data for 454 firms listed on China's STAR Market from 2019 to 2024, this paper examines whether capital market information can provide early indications of credit risk. The main findings are as follows:

First, the GARCH conditional volatility provides significant incremental predictive value. Adding GARCH conditional volatility to financial indicators and historical volatility further increases the model's out-of-sample AUC, indicating that dynamic market uncertainty can provide risk signals before financial risk becomes evident. Second, there is no simple monotonic relationship between credit risk and conditional volatility. Based on the within-quarter quintile grouping, the risk event rate rises from 3.33% in Q1 to 7.11% in Q4, with Q4 significantly exceeding Q1, while Q5 (the highest volatility group) declines to 4.62% and does not differ significantly from Q4. This pattern indicates interval heterogeneity in the risk implications of market volatility: the medium-high volatility interval reflects persistent operating uncertainty and serves as an important forward-looking signal of credit risk, whereas extreme volatility more likely reflects event-driven shocks and does not constitute a further signal of credit deterioration. Third, network centrality does not generate additional predictive gains on its own. Part of the information contained in network linkages may already be reflected in the firm's own market volatility. Fourth, information disclosure behavior provides particularly strong incremental predictive information. R&D investment, cash flow quality, operating efficiency, and consecutive gaps in R&D information emerge as important signals in the identification of credit risk.

These findings have three main implications for financial risk monitoring. First, regulators and financial institutions could incorporate market volatility and abnormal trading activity into conventional financial risk-monitoring systems. Volatility should not, however, be treated as a single linear early-warning threshold. Its interpretation should take account of the firm's financial condition and stage of innovation. Second, the continuity of innovation-related disclosure should be considered in the credit assessment of sci-tech firms. Firms with persistent gaps in key R&D information may warrant closer risk review. Third, financial institutions should shift from static financial scoring to dynamic, probability-based early-warning systems that integrate market information, innovation characteristics, and information disclosure behavior, thus improving their ability to detect the gradual deterioration of credit risk.

This paper also has several limitations. The analysis relies primarily on structured financial and market data, and has not fully utilized textual information such as announcements, news, and analyst expectations. The informational overlap between network variables and conditional volatility also limits the identification of specific risk transmission channels. Future research could integrate unstructured information and supply chain, ownership, and financial networks.

Declarations

Data availability

All data generated or analysed during this study are included in the Supplementary Information fles.

Funding

This work was funded by Social Science Research Foundation of Ministry of Education of China (15YJA790051), National Social Science Fund Project of China (17BGL058) and Natural Science Foundation of Shandong Province(ZR2022MG045), Special Project of Shandong Social Science Planning Program (23CLYJ35).

Author Contributions

Formal analysis, software, conceptualization, methodology, and writing-original draft, X.C.; supervision, review and editing, C.S. and X.S. All authors reviewed and approved the final manuscript.

Ethical Approval and Informed Consent

This article does not contain any studies with human participants performed by any of the authors.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

References

1. Altman EI (1968) Financial ratios, discriminant analysis and the prediction of corporate bankruptcy[J]. J Finance 23(4):589–609
2. Ohlson JA (1980) Financial ratios and the probabilistic prediction of bankruptcy[J]. J Accounting Res 18(1):109–131
3. Shumway T (2001) Forecasting bankruptcy more accurately: A simple hazard model[J]. J Bus 74(1):101–124
4. Merton RC (1974) On the pricing of corporate debt: The risk structure of interest rates[J]. J Finance 29(2):449–470

5. Campbell JY, Hilscher J, Szilagyi J (2008) In search of distress risk[J]. *J Finance* 63(6):2899–2939
6. Bharath AS, Shumway T (2008) Forecasting default with the Merton distance to default model[J]. *Rev Financial Stud* 21(3):1339–1369
7. Engle RF (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation[J]. *Econometrica* 50(4):987–1007
8. Bollerslev T (1986) Generalized autoregressive conditional heteroskedasticity[J]. *J Econ* 31(3):307–327
9. Hall BH (2002) The financing of research and development[J]. *Oxf Rev Econ Policy* 18(1):35–51
10. Brown JR, Petersen BC (2009) Why has the investment-cash flow sensitivity declined so sharply? Rising R&D and equity market developments[J]. *J Banking Finance* 33(10):1789–1802
11. Hall BH, Lerner J (2010) The financing of R&D and innovation[M]//Hall. In: Rosenberg BH N (ed) *Handbook of the Economics of Innovation*. Elsevier, Amsterdam, pp 609–639
12. Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions[C]//*Advances in Neural Information Processing Systems*. 30:4765–4774
13. Baker M, Wurgler J (2006) Investor sentiment and the cross-section of stock returns[J]. *J Finance* 61(4):1645–1680
14. Allen F, Gale D (2000) Financial contagion[J]. *J Polit Econ* 108(1):1–33
15. Spence M (1973) Job market signaling[J]. *Q J Econ* 87(3):355–374
16. Verrecchia RE (1983) Discretionary disclosure[J]. *J Account Econ* 5:179–194

Figures

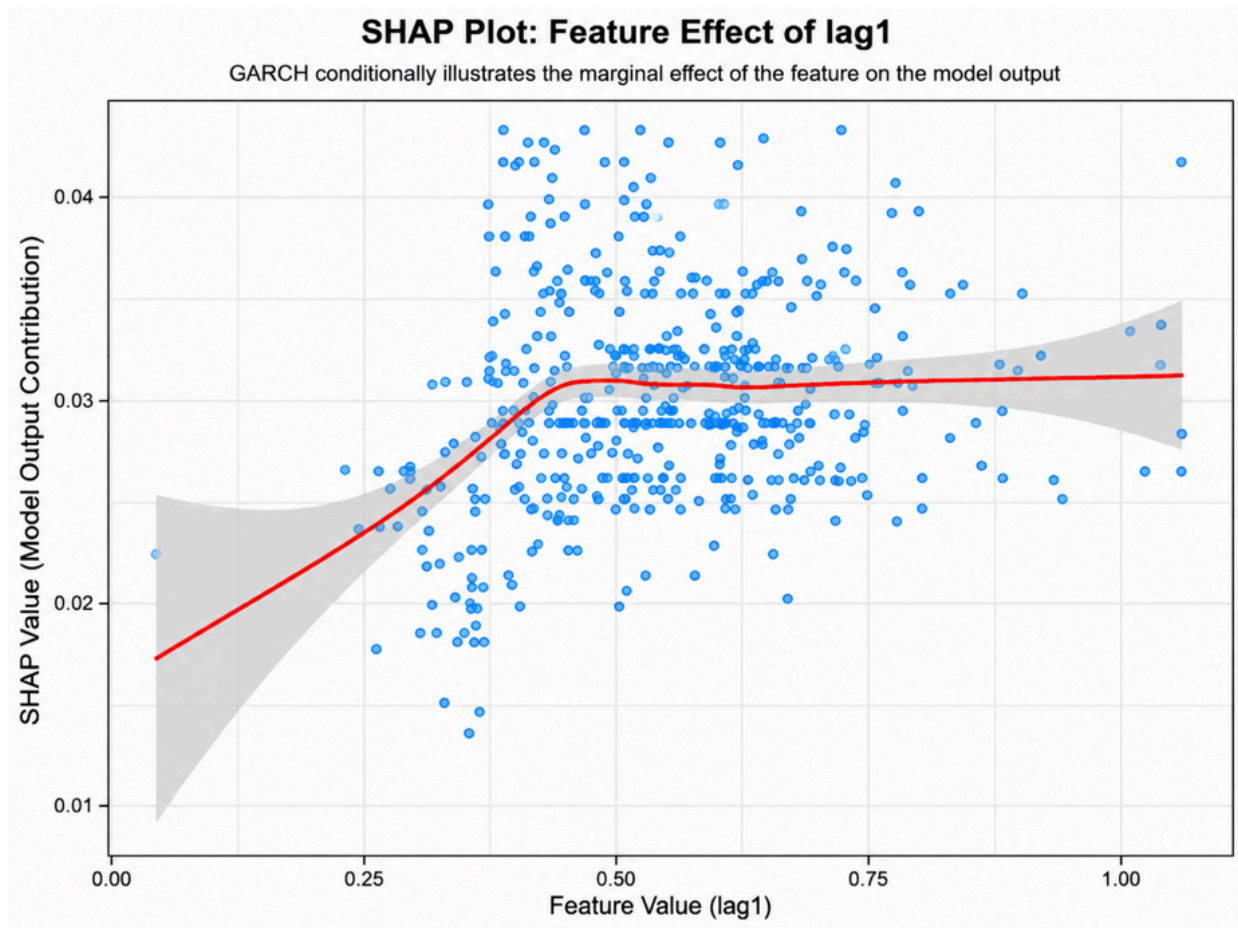


Figure 1

SHAP Interaction Dependence Plot

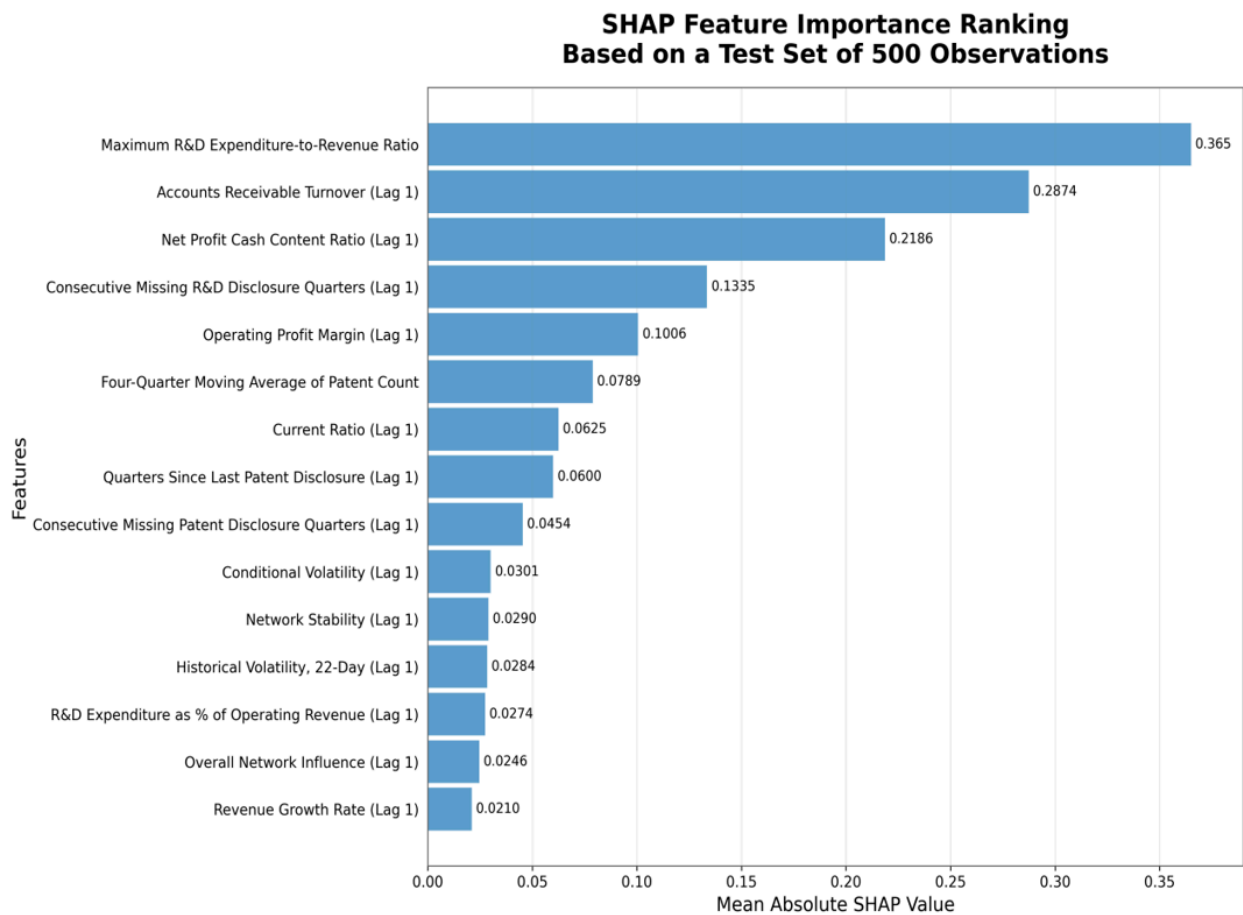


Figure 2

SHAP-Based Variable Importance Ranking

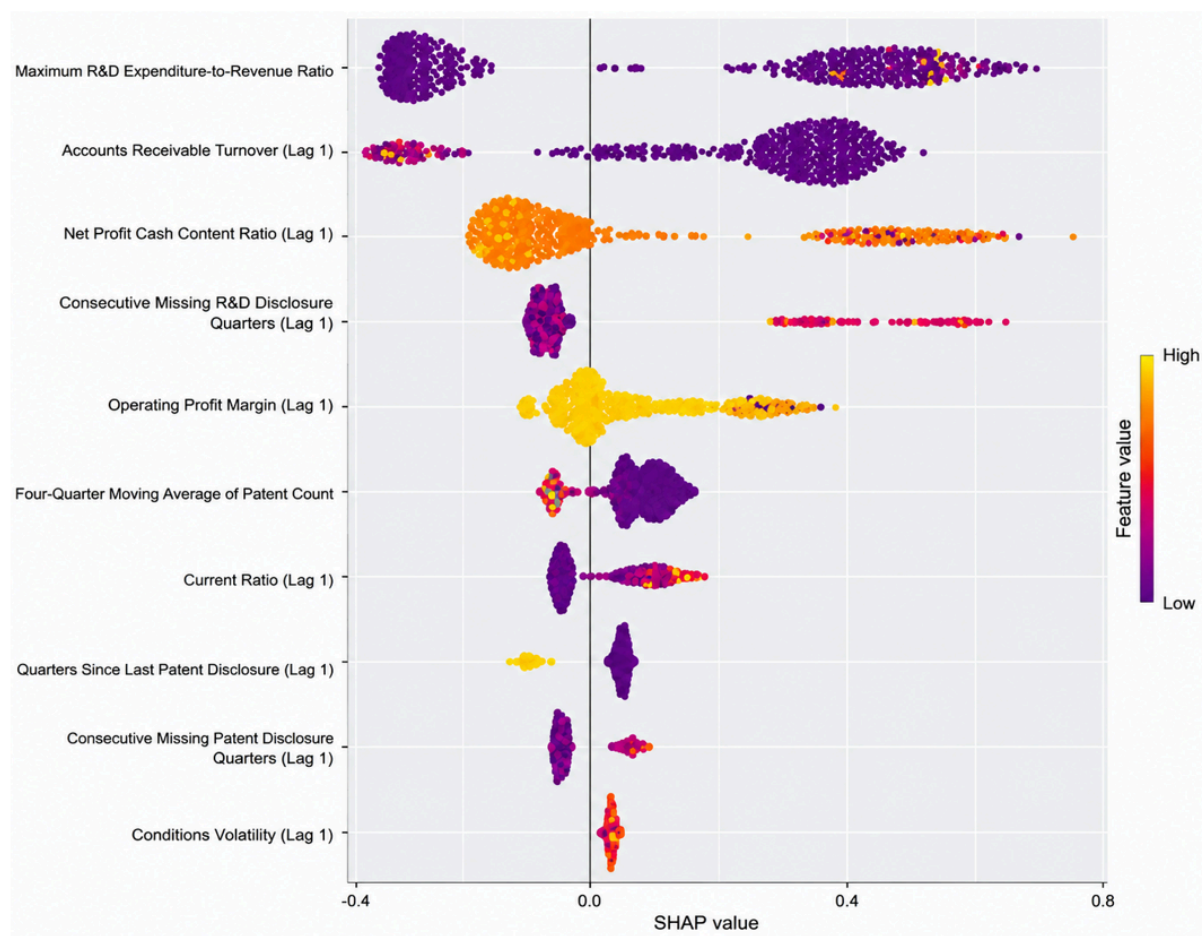


Figure 3

SHAP Summary Plot

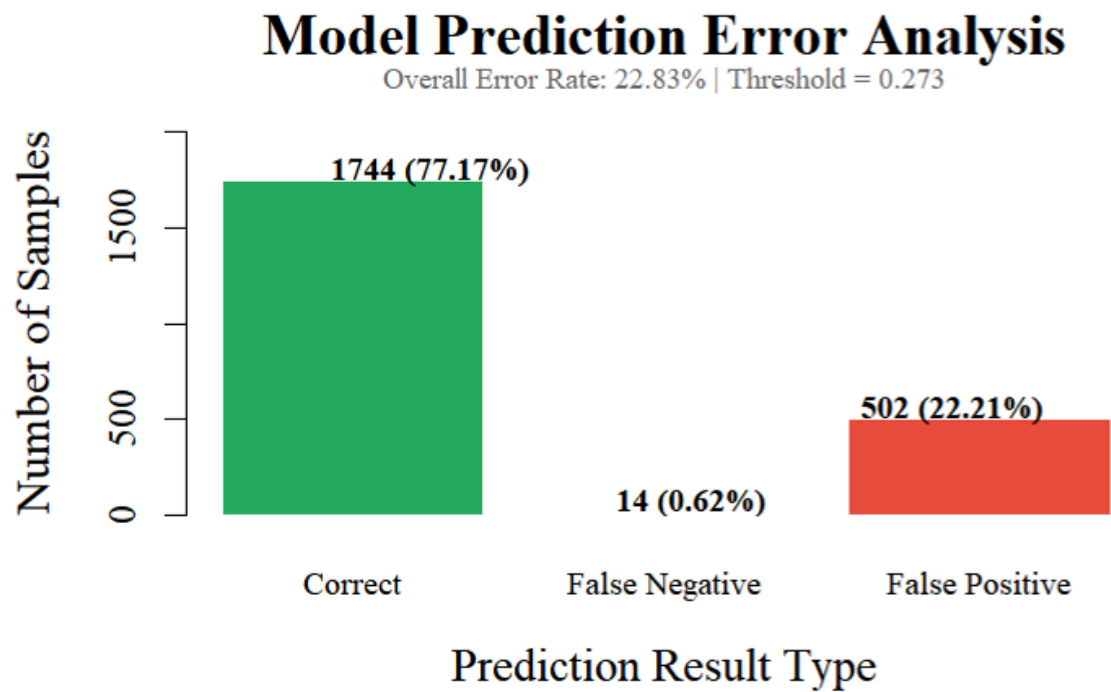


Figure 4

Model Prediction Error Analysis