frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 27 March 2025
DOI 10.3389/fphy.2025.1559283

OPEN ACCESS

EDITED BY
Ze Wang,
Capital Normal University, China

REVIEWED BY
Chan Liu,
China University of Mining and Technology, Beijing, China
Xiaowen Shi,
Beijing Normal University, China

*CORRESPONDENCE
Wei Cui,
cuiw@cugb.edu.cn

RECEIVED 12 January 2025
ACCEPTED 13 March 2025
PUBLISHED 27 March 2025

CITATION
Zhang Y and Cui W (2025) Research on characterization and prediction of bond risk factors based on machine learning: evidence from the China.
Front. Phys. 13:1559283.
doi: 10.3389/fphy.2025.1559283

Research on characterization and prediction of bond risk factors based on machine learning: evidence from the China

Yaowen Zhang and Wei Cui*

School of Economics and Management, China University of Geosciences, Beijing, China

Introduction: The scale of default on credit bonds in China has been expanding. Credit bond defaults not only increase the financing costs of enterprises but also affect the efficiency of debt issuance and even lead to the spread of risks in the financial market. Accurately identifying bond default risks, clarifying the characteristics of bond defaults, and understanding the default risk mechanism are of crucial importance.

Methods: This paper takes corporate credit bonds as the research object and analyzes bond defaults from both macro and micro perspectives. From a macro perspective, it confirms the logical transmission between macro factors and bond defaults through causal relationships and grasps the overall characteristics of bond defaults by combining association rule mining and descriptive statistical research methods. Bonds are divided into a risk-free bond group and a risky bond group, and association rules are mined in four dimensions: the bond issuance region of the enterprise, whether the issuer is listed, the attributes of the issuing enterprise, and whether the enterprise bond is guaranteed. Based on these rules, a cross-analysis of bond risk factors is conducted. From a micro perspective, taking each bond as the research object, a bond default identification system is established, and default predictions are made based on the ensemble learning algorithm. The important characteristics of default bonds are analyzed from the perspective of whether the issuer is a state-owned enterprise, and further cause difference analysis is conducted.

Results: The results show that M1 and M2 have an impact on bond defaults, and the ensemble machine learning algorithm can accurately predict bond default risks and obtain key factors for bond risk identification. It is reasonable to choose macro indicators to predict bond defaults.

Discussion: Based on the experimental conclusions, this paper discusses and analyzes the bond risk evolution process and the reasons for risk concentration in certain industries, which is helpful for a comprehensive understanding of bond default risks. Our research can provide tool references and guidance for risk management in the actual bond market.

KEYWORDS

Chinese bond market, bond credit default prediction, macro factors and default causality, combinatorial machine learning, granger causality test, association rule mining

1 Introduction

The bond market is an important component of the capital market. The effective operation of the bond market not only provides financing convenience for market participants but also promptly reflects the operational status of the real economy.

Chinese corporate bond issuance, which began in 1984, started late but has grown rapidly to become the world's second largest bond market. Investors know little about the bond market compared to other forms of financing. The bond market is very sound and low risk for most investors. For a long time, Chinese credit bonds have been regarded as risk-free investment tools by investors, but with the improvement of the marketization degree of Chinese bonds, bond defaults began to appear, and the first case of corporate bond defaults occurred in 2014. Chinese corporate bonds have had a 10-year default history since November 2014, when the first case of "Chaori bonds" actually defaulted. At present, defaults are gradually becoming normal, credit risks are spreading across industries, and market players have stopped unreasonable "risk-free" expectations for Chinese corporate bonds.

By the end of 2023, the total value of outstanding bonds in China's bond market had grown to 155.75 trillion yuan. Starting from 2018, the market witnessed a significant rise in bond defaults. The year 2019 marked the highest number of defaults, with 207 bonds failing to meet their obligations, totaling 166.187 billion yuan. Although the number of defaults dropped slightly in 2020, the total value of defaulted bonds hit its peak at 188.462 billion yuan that year. Following this period, while the overall scale of defaults has shown some reduction, even high-credit-rated companies and centrally-administered state-owned enterprises—typically considered lower-risk entities—have also faced default incidents. This development is likely to have a more profound impact on the credit bond market environment. Bond default will lead to capital loss, market risk intensification, credit risk spread, deterioration of credit environment, reduce investors' investment participation in the bond market, and even lead to systemic financial risk. Therefore, it is very important to analyze and forecast the credit risk factors of bonds. China's corporate bond defaults deserve attention.

As for the prediction methods of bond default, the earliest statistical and quantitative methods are Z-score model [1] and KMV model [2], etc. The statistical and quantitative methods of bond default are developed on the improvement of these models. In recent years, with the development of big data, machine learning algorithms have been applied in the field of credit prediction. Bao et al. [3] found that compared with the majority of existing studies that mainly aim to explain sample fraud and emphasize causal reasoning, the predictive model constructed by ensemble learning can predict accounting fraud more accurately. In the establishment of the bond default system, the current research generally selects the macro environment of the enterprise directly as the predictive index of the machine learning algorithm, while ignoring the causal relationship between the selection of macro indicators and bond default. Therefore, this study integrates causal inference with machine learning techniques to systematically analyze the causal relationships between macroeconomic indicators and bond defaults. It employs machine learning algorithms for predicting bond default risks and validates the effectiveness of the selected macroeconomic indicators through feature importance analysis.

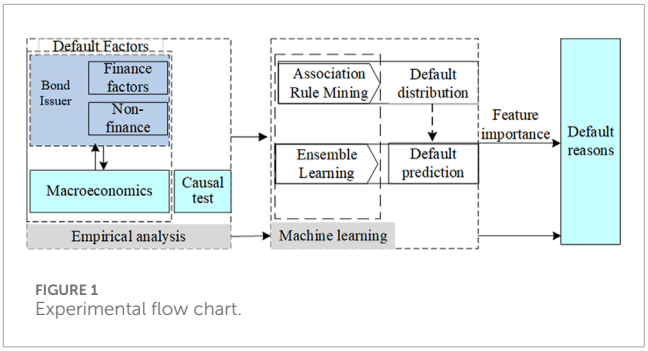

FIGURE 1
Experimental flow chart.

TABLE 1 Confusion matrix.

Actual	Prediction	
	0 (non-risk)	1 (risk)
0 (non-risk)	TN	FP
1 (risk)	FN	TP

The bond market is a crucial link in the transmission of monetary policy. China is in the process of interest rate liberalization. Grasping the risks in China's bond market and clarifying the causal relationship between monetary and financial variables, macroeconomic variables and bond defaults is of great significance.

This study focuses on credit bonds issued by Chinese companies from 2014 to 2022 as the research sample, encompassing corporate bonds, enterprise bonds, medium-term notes, and short-term financing bonds. Bonds that are in default or have been extended are regarded as risky bonds, while bonds that mature normally are considered risk-free bonds. By using causal inference to analyze the causal relationship between macroeconomic indicators and bond defaults, and at the same time providing a reference for the selection of the timing of predictive indicators, machine learning techniques are utilized to predict bond default risks, and the effectiveness of the selected macroeconomic indicators is verified through feature importance analysis. The main experiments are as follows [1]: Conduct an empirical analysis on the impact of macroeconomic factors such as macro monetary policies and the economy on bond defaults through causal inference to explore whether there is a causal relationship between macroeconomic factors and bond defaults [2]. On the basis of descriptive statistics and association rule mining of the attributes of bonds themselves, find the associated features of bond defaults [3]. Process and construct the risk indicator system, and build a bond default risk early warning model through machine learning algorithms. The main innovations and contributions of this paper are as follows [1]: In the selection of indicators, it empirically proves the causal relationship between macroeconomic indicators and bond default [2]; Association rule mining is applied to analyze the relevant characteristics of risk bonds, and the distribution characteristics and rules of risk bonds are obtained [3]; Based on the

TABLE 2 Optimal parameters.

Algorithm	Parameters	Range of values	Optimal parameters
Decision Tree	Criterion	Gini, entropy	Gini
	Max_depth	3,5,10,15,20,25,30	15
	Min_samples_leaf	[1,2,3,4,5]	1
	Min_samples_split	[2,4,6,8,10]	2
RandomForest	N_estimators	[50,1000], step = 1	71
	Max_depth	3,6,9,12,15,18,21,24,27,30	18
GradientBoostingClassifier	Learning_rate	[0.0001-0.1]	0.1
	N_estimators	[50,1000], step = 50	150
	Max_depth	[3,30], step = 1	21
	Max_features	[10,50], step = 1	33
	Subsample	[0.1,0.9], step = 0.1	0.8
	Min_samples_leaf	[60,100], step = 1	70
	Min_samples_split	[50,200], step = 1	100
Catboost	Learning_rate	[0.0001-0.05], step = 0.001	0.136
	Rsm	[0.6, 1.0], step = 0.1	0.60
	Subsample	[0.6, 1.0], step = 0.001	0.736
	Iterations	[20, 50],step = 1	50
	Depth	[3,6], step = 1	4

results of association rules, the combination learning model is used to make group prediction and factor analysis of bond credit status, so as to grasp the differences, and prove that the prediction accuracy of combination learning on bond default is improved compared with the single classification model.

2 Literature reviews

Bond risk identification is the prediction and assessment of the credit status of bond issuance. The following reviews the existing research results from three aspects: the selection of indicators for bond default prediction, including financial and non-financial indicators; the risk transmission between monetary policy and financial markets and causal analysis and machine learning.

In terms of the selection of financial indicators, in 1968, the American economist Beaver [4] was the first to use financial analysis indicators to predict the default of corporate bonds and credit, and to create a univariate financial early warning model with the idea of regression analysis. Altman [5] established a default probability prediction model to predict the default probability of enterprises according to the ratio of retained earnings to total assets and other financial indicators. Douglas A et al. [6] found that net cash flows from operating activities, investing activities and financing activities are important sources for enterprises to repay debts. Sun [7] selected eight parts such as financial structure, growth level and profitability when identifying the financial distress of enterprises.

Beaver [8] argued that relying solely on financial factors to study bond default risk is inadequate due to challenges in obtaining accurate and reliable financial information. The research by Sadiq et al. [9] further suggests that the macroeconomic environment significantly influences corporate credit risk. In terms of the selection of non-financial indicators, Chava and Jarrow [10] found industry factors can affect the operation, profitability and development of enterprises in the same industry and industry cycle, and thus lead to bond default. Wu [11] selected three levels of non-financial indicators: macro, industry and enterprise. Eugene [12] and Wilson [13] discovered that when the macroeconomic environment is favorable, *per capita* disposable income grows more rapidly, which in turn leads to an increase in investment expenditure. The demand for bond investment rises, and enterprises have a stronger ability to repay principal and interest on bonds, with a lower risk of default. Lu Jun et al. [14] found that macroeconomic indicators such as GDP, CPI and M1 growth rate of narrow money supply had a significant

TABLE 3 Results of descriptive statistics.

Index	Mean	Std	Min	25%	50%	75%	Max
M1	518184.46	110831.42	327220.2	435628.1	543867.4	607266.8	674374.8
M2	1830854.03	445101.38	1,160,687	1,479,418	1,785,922	2,169,763	2,664,321
Business prosperity index	120.95	10.13	83.66	114.3	122.3	126.9	145.9
EPU	5.43	0.52	4.33	4.93	5.58	5.78	6.21
GDP	106.68	2.35	94.6	106.1	107	108	112.85
CPI	101.93	0.59	100.9	101.6	102	102.1	102.9
Local government debt	42.67	28.47	0	27	37	54	190
Proportion of shares held by major shareholders	64.63	30.42	4.03	38	64.84	100	100
Return on equity	3.29	33.99	−2204.89	1.2	4.25	8.54	183.86
EBIT/Total operating revenue	20.91	191.60	−4169.95	4.6	10.36	21.88	16185.02
Sales expenses/Total operating revenue	4.27	7.92	−2.27	0.79	2.35	4.96	336.7
Financial expenses/Total operating revenue	8.55	74.61	−1047.62	1.14	2.935	6.54	4725.95
Asset impairment loss/Total operating revenue	2.10	28.55	−431.47	−0.1	0.05	0.66	1534.54
Net income from operating activities/Total profit	−2.13	7186.05	−174767	23.1975	71.375	91.9525	650990.3
Net non-operating income and expenses/Total profit	49.67	1320.35	−4147.21	−0.13	1.15	8.425	112496.3
Income tax/Total profit	40.09	408.15	−2395.67	16.43	23.35	28.31	27516.8
Cash received from sales of goods and provision of services/Operating revenue	537.70	30087.72	−16.6	94.59	105.38	114.62	2,902,529
Long-term asset suitability rate	927857.00	89,754,481.00	−61399.8	133.2275	217.1	477.74	9.13E+09
Current assets/Total assets	47.90	23.62	0.61	29.2475	46.715	65.305	100
Current liabilities/Total liabilities	77.14	1108.30	0.62	51.83	67.86	81.24	109084.6
Quick ratio	1.08	7.68	0.02	0.55	0.79	1.1	707.8
Tangible assets/Interest-bearing debt	299.90	21570.23	−6.75	0.215	0.52	1.09	1,866,983
Net cash flows from operating activities/Non-current liabilities	−163.22	17768.32	−1268235	−0.06	7.62	26.61	219800.2
Tangible net worth debt ratio	−91.80	37889.14	−2971202	141.42	295.91	607.64	1,077,645
Inventory turnover rate	85.00	2158.91	−267.72	1.19	3.35	7.8625	136849.2
Accounts receivable turnover rate	100.31	2867.55	−1195.47	3.17	6.97	17.5875	219474.1
Accounts payable turnover rate	9435.04	675545.30	−10.08	1.97	3.81	7.985	58,719,863
Cash turnover rate	1053.23	12497.48	−5701.92	188.9	441.92	895.275	748911.4
Total operating revenue (year-on-year growth rate)	29.37	642.10	−100	−3.485	8.435	23.29	57909.39
Net profit (year-on-year growth rate)	−29.42	2996.55	−164908	−24.925	9.01	45.125	87598.59
Net assets (year-on-year growth rate)	9.30	115.05	−7951.27	1.38	7.33	16.51	1302.31
Total assets (year-on-year growth rate)	15.10	80.46	−2804.9	2.2175	9.45	19.52	4819.76

(Continued on the following page)

TABLE 3 (*Continued*) Results of descriptive statistics.

Index	Mean	Std	Min	25%	50%	75%	Max
Net cash flow (year-on-year growth rate)	523.87	19662.41	−112345	−122.9	22.01	165.65	1,554,680
Cash and cash equivalents/Short-term debt	33.33	1884.41	0	0.3	0.56	1.12	164211.5
Proportion of cash flows from financing activities	−1832.54	192496.20	−1.30E+07	−95.12	65.69	243.29	4,024,864
Proportion of cash flows from investing activities	1228.81	94828.06	−1701245	−170.4	−5.24	124.985	6,091,123
Proportion of cash flows from operating activities	703.52	124983.90	−4570310	−118.4	38.67	212.01	7,347,279

impact on companies falling into financial crisis. Chen [15] studies the relationship between macroeconomic environment and credit risk; At the macro level, LIU [16] selected CPI, social financing scale, market interest rate and so on.

Bonds belong to corporates' debt; debt repayment ability is closely related to the liquidity of enterprises. Money supply is a measure of market liquidity, it has an important impact on the price of various assets in the financial market, and its change will directly affect the economic environment. China's infrastructure projects account for a large share of the economy, money supply tends to outpace economic data. Fan et al. [17] studied the relationship between the excess rate of return of bonds and macroeconomic variables, and believed bond prices and market interest rates were mutually determined, and market interest rates were affected by factors such as money supply M1 and actual consumption. Through empirical evidence, they proved that changes in market interest rates might affect the expected excess rate of return of bonds. Based on the Merton structured model and the data of China from 2000 to 2010. Dai and Sun [18] studied the factors affecting the credit spread of corporate bonds in China's Shanghai and Shenzhen bond markets, and came to the conclusion that the issuance of M1 had a positive impact on the credit spread of corporate bonds. The fluctuation of stock index can indicate the current macroeconomic situation of economy and society to a certain extent, and the major changes of economic data will be reflected in the stock index. The stock market index is the vane of the stock price market, so the stock market index can also indicate the macro-economic situation.

In the selection of machine learning metrics, the causal relationship between metrics is a problem that should be given due attention. Some scholars have attempted to verify the relationship between the indicators and the predicted targets through causal relationship tests. In the field of transportation, Zhang et al. [19] studied the traffic speed of taxi in Changsha, China's urban road network, and tested the spatiotemporal causality between various links through nonlinear Granger causality, providing reliable guidance for MTL model to select information features from the whole network link. Luo [20] applied nonlinear Granger causality test to explore the causal relationship between traffic areas, built a multi-task deep learning model framework with long short-term memory (LSTM) as the core neural unit, and verified it with New York City taxi trip data. In the field of economics and finance, Wei et al. [21] selected the three variables with the highest correlation with the exchange rate among 16 macroeconomic variables including import, export and foreign exchange reserves through Granger causality test analysis, and used KELM to make a medium-term forecast of RMB against US dollar. The results show that the set learning method based on KELM outperforms all other benchmark models in different prediction periods. Xu et al. [22] combined Granger causality test with support vector machine and found that among the 15 stocks studied, the after-hours collective sentiment of nine stocks had a strong predictive effect on stock price changes the next day. In the field of medicine, Almalaq et al. [23] use Granger causality test to determine the activation region related to the verbal fluency task in human EEG, and train the classifier between subjects with Parkinson's disease and healthy control group in combination with support vector machine (SVM). In the field of physics, H et al. [24] introduced Granger causality to analyze the relationship between sensor variables, and selected variables with greater Granger causality relationship with sensor loss data as the input vector of the extreme learning machine. In the field of ecological environment, Vazquez-Patino et al. [25] explored the improvements in interpretability and robustness of models using causal selection predictors, comparing three cause-based methods with ML's four standard predictor selection methods using rainfall data from the Andean basin in Ecuador. Li et al. [26] screened the key environmental factors affecting sea ice concentration based on causal relationship (KGC), and used a variety of machine learning (ML) algorithms to make short-term sea ice prediction.

The literature reviewed highlights the related factors of bond risks and the application of causal analysis in machine learning. It is evident that scholars across various fields emphasize the importance of the correlation between selected indicators and the target when applying machine learning and deep learning algorithms. Through an analysis of the factors influencing corporate bond defaults, it has been found that key factors related to identifying corporate bond credit status include the financial metrics of the enterprise itself, which are the most representative indicators of its operational conditions. Non-financial indicators, such as macroeconomic factors, have also garnered attention from relevant scholars. Based on these findings, this paper employs causal analysis to select critical predictive indicators in the field of bond risk identification to validate the effectiveness of the chosen macroeconomic factors. Additionally, association mining is incorporated to describe the characteristics of risk aggregation. On this foundation, financial data and associated macroeconomic factors are selected, and ensemble learning algorithms are utilized to predict the credit status of

TABLE 4 Bond risk indicator system.

Factor	Indicator	Indicator	Indicator	Factor	Indicator
Profitability	Return on Equity (ROE)	Growth capacity	Total operating revenue (year-on-year growth rate)	Profit quality	Net income from operating activities/Total profit
	Ebit/Total operating revenue		Net assets (year-on-year growth rate)		Net non-operating income and expenses/Total profit
	Sales expenses/Total operating revenue		Net cash flow (year-on-year growth rate)		Income tax/Total profit
	Asset impairment loss/Total operating revenue		Net profit (year-on-year growth rate)	Local economy	GDP index
	Financial expenses/Total operating revenue		Total assets (year-on-year growth rate)		Local government debt
Cash flow	Net cash flow from operating activities ratio	Bond	Bond issuance rating	Corporate governance	Shareholding ratio of shareholders
	Net cash flow from financing activities ratio				Company type
	Net cash flow from investing activities ratio		Rating before maturity (default)	Macroeconomic	CPI, M1, M2, EPU
	Cash received from sales of goods Provision of services/Operating revenue		Rating change		
Capital structure	Tangible assets/Interest-bearing debt	Debt-paying ability	Quick ratio	Operating capacity	Inventory turnover rate
	Long-term asset suitability rate		Net cash flow from operating activities/non-current liabilities		Accounts receivable turnover rate
	Current assets/Total assets		Tangible net worth debt ratio		Cash turnover rate
	Current liabilities/Total liabilities		Cash And Cash Equivalents/Short-term debt		Accounts payable turnover rate
Industry impact	Business prosperity index				

specific bonds issued by enterprises, thereby enhancing the accuracy of identification. The main contributions of this paper lie in the following aspects: In terms of algorithm selection, traditional methods (such as logistic regression and probit models) perform poorly in handling nonlinear relationships and complex data structures. They have limited capacity to process high-dimensional data and are unable to capture the interactions among variables. We employed multiple ensemble machine learning algorithms and selected the optimal model through comparative analysis, significantly improving the prediction accuracy. In the construction of the indicator system, previous studies typically used only a single data source, lacking the integration of multi-source data. We developed a multi-dimensional bond default identification system that integrates macroeconomic indicators and market data, providing more comprehensive data support. Regarding the selection of macro indicators, previous studies often directly

TABLE 5 Stability test.

Variables	Original sequence	First difference sequence
Default	Non-stationary time series	stationary time series
M1	Non-stationary time series	stationary time series
M2	Non-stationary time series	stationary time series
HS300	Non-stationary time series	stationary time series
CPI	Non-stationary time series	stationary time series

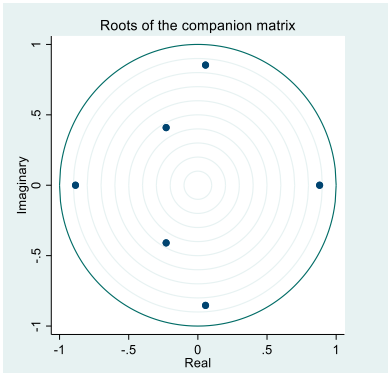

FIGURE 2
VAR stability test.

TABLE 6 ADF test results.

Variables	ADF	P-value	Critical value 1%
d.lnM1	−9.751	0	−3.689
d.lnM2	−6.169	0	−3.689
d.lnDefault	−8.374	0	−4.297

incorporated macro indicators into the prediction model without clarifying their relationship with bond defaults. This paper aims to demonstrate the causal relationship between macroeconomic indicators and bond defaults.

3 Research design and methodology

This article conducts bond risk identification. From a macro perspective, We use Granger causality test to analyze the causal relationship between macroeconomic indicators and the amount of bond default, and association rule mining is used to grasp the risk characteristics of bonds. From a micro perspective, individual bond samples are taken as the research objects. Combined with combinational learning classification algorithm, credit default problems of corporate bonds issued by companies in China are predicted. We select the data of corporate credit bonds issued by Chinese companies from 2014 to 2022. Including short-term financing bills, corporate bonds, enterprise bonds, medium-term notes, and asset-backed securities The main research objectives are divided into three parts: First, for the selected macroeconomic factors, we apply co-integration test and Granger test to verify the causal relationship between macroeconomic and bond default; Secondly, since most machine learning algorithms have limited interpretation of text type indicators, we use association rule mining algorithm to carry out visual cross-analysis of the characteristics of defaulted bonds, and grasp the distribution of default risk from a macro perspective. Finally, based on the verified results, we use the combinatorial learning model to forecast corporate bond default, compare the forecast results before and after adding macro factors, and conduct in-depth analysis of default factors. (Figure 1).

3.1 Causal analysis

VAR constructs the econometric analysis model by taking each endogenous variable in the economic system as a function of the other endogenous variables and their lagging values. It can deal with the estimation problems caused by endogenous variables as explained variables, identify and estimate the interdependence of endogenous variables or interactive spillover relationships. In Formula 1, y_t is a k-dimensional endogenous variable, x_t is an exogenous variable, p is the number of lag periods, A and B respectively represent matrices composed of estimated coefficients, and ε_t represents the error term. In the context of bond default analysis, VAR is used to analyze the relationship between Default and M1, M2, Default and CPI.

$$y_t = A_1 y_{t-1} + \ldots\ldots + A_\beta y_{t-p} + Bx_t + \varepsilon_t (t \in (1, T)) \quad (1)$$

3.2 Algorithm and evaluation index

APRIOR is used for association rule mining, APRIOR is a frequently used algorithm for mining frequent itemset, which was formally proposed by Agrawal et al [27]. In 1994. It features multiple scans of the database, a large scale of candidate items, and high computational cost for support calculation. By adopting an iterative approach, the algorithm steps are divided into two parts: "connection" and "pruning".

The reasons for choosing decision trees as the base learners are as follows: Compared with linear regression and other methods, decision trees have no strict requirements for data distribution, are robust to missing and noisy data, and can automatically handle and reduce the impact of outliers. Compared with neural network algorithms, decision tree models have an intuitive structure, are easy to interpret and visualize, provide a clear decision path for research problems, and have strong interpretability. Compared with algorithms such as KNN, they can efficiently handle high-dimensional data and reduce the influence of redundant features. Ensemble learning improves model performance by combining multiple base learners. Compared with a single machine learning model, ensemble learning integrates the prediction results of multiple models to reduce the bias and variance

TABLE 7 Trace test of Default, M1 and M2.

Maximum Rank	Params	LL	Eigenvalue	Trace statistic	Critical Value 5%
0	15	136.04948		76.7476	34.55
1	20	160.25203	0.79017	28.3425	18.17
2	23	170.15279	0.47205	8.5410	3.74
3	24	174.42327	0.24082		

TABLE 8 Default, M1, M2 maximum characteristic root test.

Maximum rank	Params	LL	------Eigenvalue----- Maximum		Critical value 5%
0	15	136.04948		48.4051	23.78
1	20	160.25203	0.79017	19.8015	16.87
2	23	170.15279	0.47205	8.5410	3.74
3	24	174.42327	0.24082		

TABLE 9 D.lnDefault D.lnM1 D.lnM2 Optimal lag order.

Lag	LL	LR	DF	P	FPE	AIC	HQIC	SBIC
0	123.521	1	1	1	3.7e-08	−8.60862	−8.56498	−8.46588
1	135.894	24.747	9	0.003	2.9e-08	−8.84957	−8.67503	−8.27862
2	162.614	53.441	9	0.000	8.4e-09*	−10.1153*	−9.80987*	−9.11617*
3	167.042	8.855	9	0.451	1.2e-08	−9.78871	−9.35235	−8.36135
4	179.946	25.808*	9	0.002	1.1e-08	−10.0676	−9.5003	−8.212

*Optimal lag.

of a single model, thereby improving overall prediction accuracy. By majority voting or weighted averaging, it reduces the influence of data noise, thus having stronger robustness to noisy data and outliers. A single model may overfit the training data, while ensemble learning achieves a balance between model complexity and generalization ability by integrating the prediction results of multiple models. The research purpose of this paper is to identify the key factors of bond default and find an algorithm that can accurately predict bond default in practice. Given the robustness, interpretability, efficiency, and flexibility of decision tree ensemble algorithms, we choose the decision tree-based ensemble learning algorithm to predict bond risk.

Bagging and Boosting combined learning algorithms are used. Bagging is a homogeneous estimator composed of many decision trees. Boosting base learner construction has a sequence [28]. Bond risk prediction employs decision trees, the bagging algorithm based on decision trees - random forest, and boosting algorithms based on decision trees, including Xgboost, Catboost, GBDT and Hist-GBDT. GBDT (Gradient Boosting Decision Tree), also known as MART (Multiple Additive Regression Tree), The concept consists of Regression Decision Tree (DT), Gradient Boosting (GB) and Shrinkage. GBDT constructs a set of weak learners that add up the results of multiple decision trees as the final predictive output. GBDT iterates the model through gradient descent to reduce the impact of the cost function. By calculating the negative gradient of the loss function to construct residuals, it gradually reduces the value of the loss function, thereby obtaining increasingly accurate learners. GBDT can be expressed in the Equation 2. Xgboost consists of two parts: empirical risk and structural risk (regularization term), and it uses the forward stagewise algorithm to gradually optimize the classifier. Xgboost can be expressed as Equation 3. By effectively processing categorical features and introducing ranking boosting strategies, CatBoost addresses gradient bias and prediction shift issues [29]. CatBoost can efficiently handle class quantities, prevent overfitting, and train models with high accuracy. Its built-in algorithm can automatically transform categorical features into numerical features, and combine features according to the intrinsic relationship of features

TABLE 10 M1, M2 and Default Granger causality test.

Equation	Excluded	chi2	Df	Prob > chi2
D_lnDefault	D.lnM1	5.0613	2	0.08
D_lnDefault	D.lnM2	0.2828	2	0.868
D_lnDefault	ALL	6.3535	4	0.174
D_lnM1	D.lnDefault	6.8288	2	0.033
D_lnM1	D.lnM2	2.9831	2	0.225
D_lnM1	ALL	12.35	4	0.015
D_lnM2	D.lnDefault	5.3539	2	0.069
D_lnM2	D.lnM1	29.059	2	0
D_lnM2	ALL	34.044	4	0

to enrich the feature dimension. Its biggest characteristic lies in its ranking idea. HistGradientBoostingClassifier uses histogram data structure to arrange data samples implicitly, and only the largest split nodes are considered in the tree building process, so the number of split nodes is small, and only the initial input data needs to be sorted. Most parts of the HistGradientBoostingClassifier algorithm are implemented in parallel, which can effectively improve the construction efficiency of base classifier.

$$F(x) = F_M(x) + \sum_{m=1}^{M} \eta \sum_{j=1}^{J} \gamma_{jm} \Pi\left(x \in R_{jm}\right) \tag{2}$$

$$Obj = \sum_{i=1}^{N} L|F_m(x_i), y_i| + \sum_{j=1}^{N} \Omega\left(f_j\right) = \sum_{i=1}^{N} L|F_{m-1}(x_i) + f_m(x_i), y_i| \tag{3}$$

Common evaluation metrics in machine learning include confusion matrices and AUC curves. Bond risk assessment is a binary classification problem, so a confusion matrix can be used to represent the four possible outcomes of model operation, namely, FP, TP, FN, and TN, as shown in Table 1. Through the confusion matrix, accuracy, precision, recall, and F-score can be calculated.

$$Accurary = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

$$\text{Pr}ecision = \frac{TP}{TP + FP} \tag{5}$$

$$\text{Re}call = \frac{TP}{TP + FN} \tag{6}$$

$$FPR = \frac{FP}{FP + TN} \tag{7}$$

$$F_{Score} = \frac{2\ \text{Pr}ecision * \text{Re}call}{\text{Pr}ecision + \text{Re}call} \tag{8}$$

Accuracy is the ratio of the number of correctly predicted samples to the total number of samples (Formula 4). Precision is the ratio of the number of samples predicted as positive and actually positive to the total number of samples predicted as positive (Formula 5). Recall is the ratio of the number of samples that are truly positive and correctly predicted as positive to the total number of samples that are actually positive (Formula 6). False positive rate is the ratio of the number of samples predicted as positive but actually negative to the total number of samples that are actually negative (Formula 7). The F1 score is the harmonic mean of precision and recall (Formula 8). The horizontal axis of the ROC curve is the false positive rate, and the vertical axis is the recall rate. The area under the ROC curve is denoted as AUC. Generally, when AUC >0.5, the model results are considered meaningful. Recall and AUC are used to evaluate the accuracy of classification models. Before conducting machine learning experiments, parameter optimization is required. The goal of parameter optimization is to improve the prediction accuracy by optimizing the model's hyperparameters, prevent overfitting, and enhance the model's generalization ability. We use grid search and cross-validation in Python's scikit-learn, with AUC as the evaluation metric, to determine the reasonable values of hyperparameters. By traversing the predefined parameters, the parameters with the best performance are selected. The detailed parameter adjustment results are shown in Table 2.

4 Data and analysis

The research data of this article is sourced from Wind database and the National Bureau of Statistics of China. Table 3 is the descriptive statistics of the data. Through the analysis of literature review and the sorting out of the evolution process of bond risks, the factors selected in this article include macroeconomics, bond characteristics, local economy, industry characteristics, and the financial status of bond issuers. When analyzing the financial bond market, macroeconomic indicators and market indices are important influencing factors. M1 and M2 are key indicators for measuring the money supply. The changes in M1 and M2 directly affect market liquidity and interest rates, thereby having a significant impact on the bond market. CPI is an important indicator for measuring the level of inflation. Inflation expectations directly affect bond yields and prices. Based on the previous studies by Lu et al., Fan et al., and Dai and Sun, monetary indicators and CPI can affect bond defaults, so we incorporate them into the macroeconomic dimension. The economic environment of the region where the bond issuer is located is also an important influencing factor, so we include local GDP in the local economic dimension. At the same time, the characteristics of the bond itself cannot be ignored. The financial status of the bond issuer includes six aspects: capital structure, profitability, cash flow, debt-paying ability, operational ability, and development ability. There is an inherent correlation and synchronous change among financial indicators, and the contribution of financial indicators will affect the stability, explanatory power, and predictive performance of the model. Therefore, we conducted a correlation test and deleted the correlated indicators with a correlation coefficient greater than 0.6. We used the indicators that passed the correlation test to construct a bond risk identification index system (Table 4). Based on the principles of systematisms, scientific, and feasibility of the indicators, after data processing, a total of 10,516 bond data were obtained, including 657 default samples. The training set and test set were divided in a 7:3 ratio.

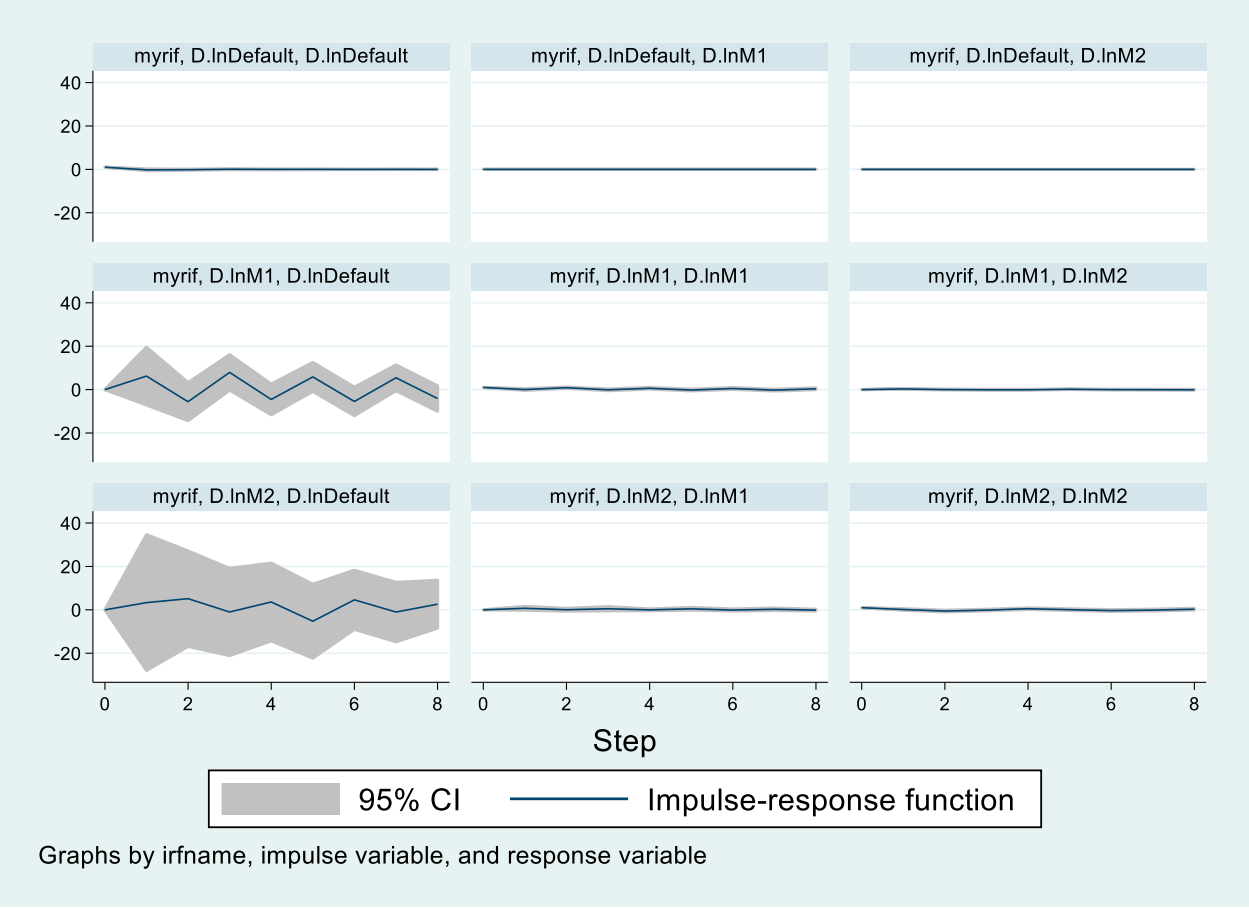

FIGURE 3
Default and M1 and M2 pulse response.

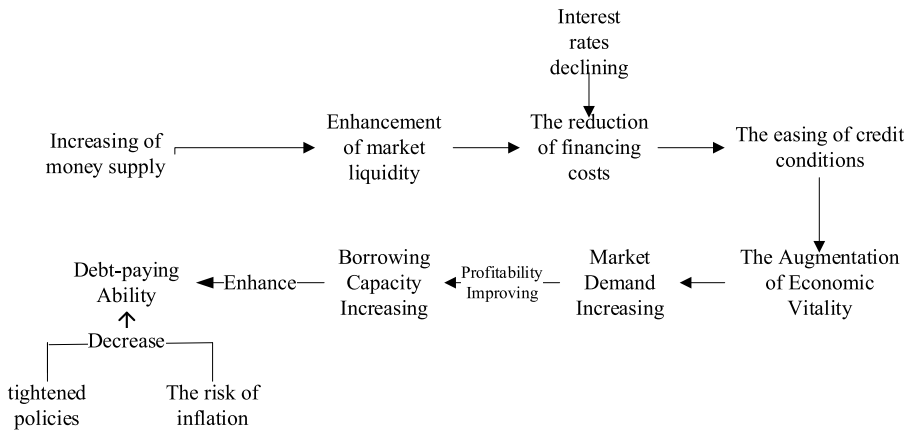

FIGURE 4
Macro risk transmission map.

4.1 Granger causality test

The purpose of the stability test is to assess whether these time series data are stationary. Stationarity tests were conducted on the default rate, M1, M2, and CPI for each quarter from 2014 to 2022. After first-order differencing, the time series of all variables became stable (Table 5). We select the reciprocal distribution of feature roots to verify the stationarity of VAR. When the reciprocal of

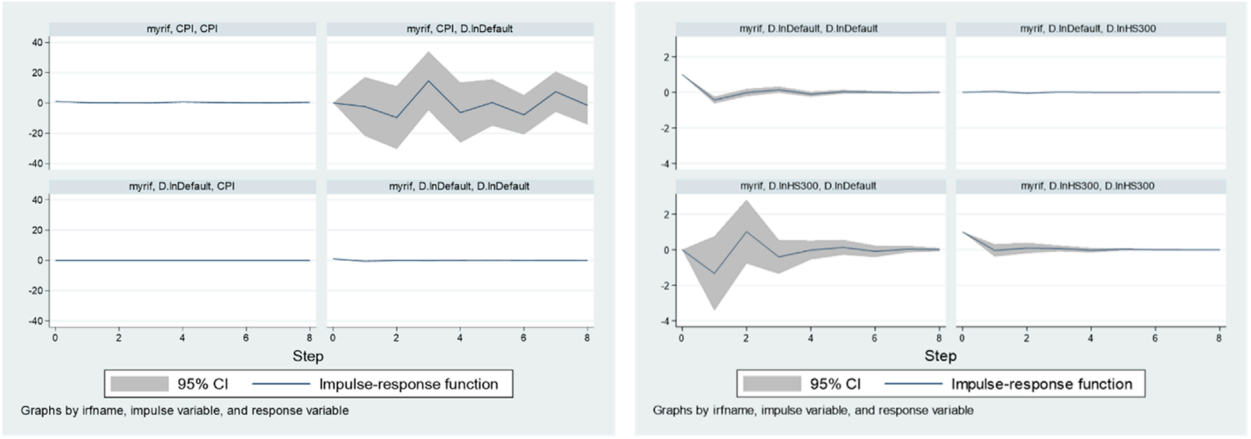

FIGURE 5
Default and CPI, Default and HS300 pulse response.

TABLE 11 Riskless bond association rules.

Support	{Local state-owned enterprises} → {unlisted companies, Riskless bonds}
Confidence	{Unsecured, Central state-owned enterprises} → {North China, Riskless bonds}
Lift	{Central State-owned enterprises} → {North China, Riskless bonds}
Frequency	Guarantee > Company Nature > Region > Industry

TABLE 12 Risk bond association rules.

Support	{Unlisted companies} → {risky bonds, unsecured} {Unsecured} → {risky bonds, unlisted companies}
Confidence	{unlisted companies, materials} → {risky bonds, unsecured}
Lift	{East China} → {Risky bonds, unsecured, unlisted companies, private enterprises}
Frequency	No guarantee, listed or not and regions are High frequency attributes

feature roots is distributed in the unit circle, the model is stable. The results show that all the reciprocal of the characteristic roots are within the unit circle (Figure 2), indicating that the model is stable. Additionally, we conducted the ADF test (Table 6), and the ADF statistics of all variables are significant, suggesting that the data series is stable. Therefore, the stability of the VAR has been fully verified. The VAR system meets the conditions for impulse response analysis.

The cointegration test of Default, M1 and M2 is carried out. The trace test results and the maximum eigen Root results show that the cointegration relationship between variables is significant (Table 7; Table 8). Table 9 shows the selection results of the optimal lag order. LR = 4, FPE, AIC, HQIC, SBIC = 2. According to the characteristics of each index, the lag order is determined as 2. The data is quarterly, which takes into account the causal relationship between factors with a lag of 6 months and defaults. According to the results of cointegration test, Granger causality test is used to explore the causal relationship between the above variables. The results are shown in Table 10. M1, M2 have an impact on bond default respectively, while M1 and M2 have an impact on the overall system respectively. Default only has an impact on M1. Therefore, there is a two-way causal relationship between M1 and the default amount of bonds, and a one-way causal relationship between M2 and the default amount. M2 has an effect on M1, while M1 has no effect on M2.

The impulse response graph is shown in Figure 3. The horizontal axis of the impulse response function represents the lag period, and the vertical axis reflects the response fluctuation level of the target variable to the unit shock. The solid line and the shaded area represent the impulse response function. Based on the corresponding impulse images, we analyzed the response process of M1 and M2 to default. The impact of M1 on default generally shows periodic fluctuations, with positive and negative effects being symmetrical. When M1 changes, default first responds positively and then moves in the opposite direction, with a period of approximately 2, that is, 1 year. After moderate symmetrical fluctuations, it tends to converge. Compared with M1, the response path of M2 to the default amount has smaller fluctuations. When the impact of M2 on default does not exceed 1 year, default first rises slowly in a positive direction and then drops in a negative direction. After that, the fluctuation period of the impact gradually shortens. Based on the above findings, we demonstrated the connection between macroeconomic monetary indicators and bond risk and sorted out the risk evolution process (Figure 4). Loose monetary policy leads to an increase in the money supply, thereby enhancing market liquidity, reducing financing costs, easing credit conditions, and gradually affecting the short-term and long-term debt repayment capabilities of enterprises; the strengthening of economic activities, the rise in asset prices, and the improvement in consumer demand will further enhance the debt repayment capabilities of enterprises. We also verified the monetary supply and credit transmission mechanism through impulse response analysis.

TABLE 13 Risk bonds strongly associated attributes.

Dimension	Value	Characteristics
Enterprise Dimension	Private enterprise	The confidence of 7 rules is higher than 95%
	listed or not	The top five rules are related to non-listed companies
Industry	Industrial sector	Industrial bonds account for a large proportion
	Material industry	Confidence of {materials}→{risky bonds, unsecured} = 97.47%
	Optional consumption	{Optional consumer sector}→{risky bonds, unsecured}
	Real estate	{Real estate}→{risky bonds, private enterprises}
Region	East China	Developed economy; Many private enterprises
	South China	
Guarantor or not	No Guarantor + Industrial	Often associated with others in association rules
	East China + Industrial	

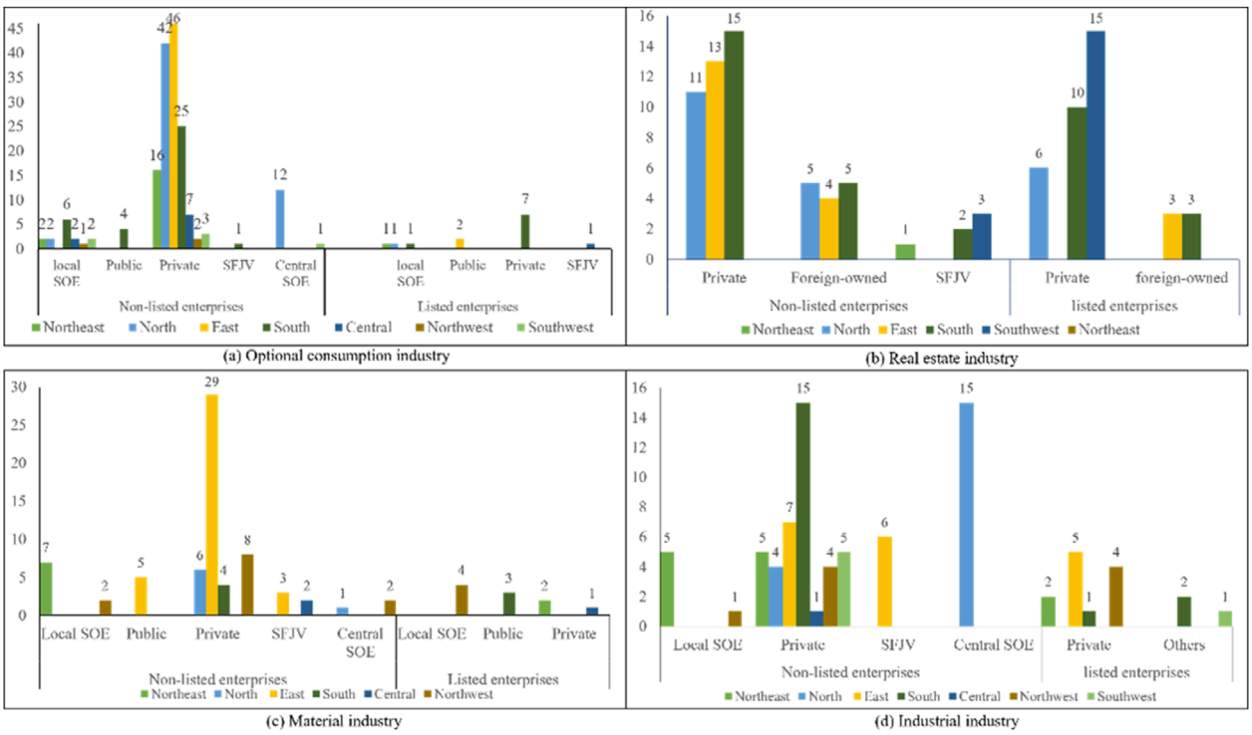

FIGURE 6
Risk distribution of default industry. **(a)** Distribution of characteristics of defaulted bonds in optional consumption industry; **(b)** Distribution of characteristics of defaulted bonds in real estate industry; **(c)** Distribution of characteristics of defaulted bonds in Material industry; **(d)** Distribution of characteristics of defaulted bonds in Industrial industry. SFJV, Sino-foreign joint venture.

Following the above steps, we explore the impact of stock market factors on bond market and the impact of CPI on bond default. The pulse response plots and results are shown in Figure 5. The impact of CPI on bond default lasts for a long time. From the overall trend, the pulse effect given by CPI firstly have a negative impact on default, and then tend to be positive. HS300 index is selected as the characteristic index of stock market. There is Granger causality between HS300 and default. Under the influence of HS300, default first shows a relatively rapid negative change and then changes into a positive change, with a cycle of about 3 quarters. After a positive and negative influence appears alternately, the influence on the default amount tends to be stable.

TABLE 14 Classification prediction results of bond risk models.

	Model		Decision	RandomForest	GBDT	Hist-GBDT	Catboost
SOEs	Accuracy	Matured bonds	0.99	0.99	1	1	1
		Risky bonds	0.50	0.82	0.49	0.85	0.84
	Recall rate	Matured bonds	0.99	1	0.99	1	1
		Risky bonds	0.65	0.68	0.85	0.85	0.94
	F1-score	Matured bonds	0.99	1	0.99	1	1
		Risky bonds	0.56	0.74	0.62	0.85	0.89
	Accuracy rate		98.36%	99.23%	98.31%	99.52%	99.61%
Non-SOEs	Accuracy	Matured bonds	0.97	0.98	0.99	0.99	0.99
		Risky bonds	0.81	0.93	0.80	0.92	0.95
	Recall rate	Matured bonds	0.96	0.99	0.96	0.99	0.99
		Risky bonds	0.86	0.88	0.95	0.95	0.95
	F1-score	Matured bonds	0.97	0.98	0.97	0.99	0.99
		Risky bonds	0.83	0.90	0.87	0.94	0.95
	Accuracy rate		94.26%	97.04%	95.47%	97.96%	98.33%

4.2 Characteristic analysis of defaulted bonds

Based on credit status, corporate bonds can be classified into matured bonds and defaulted bonds. By comparing the five attributes of the issuer's location, company characteristics, primary industry, whether there is a guarantor, and whether the issuer is a listed company, we analyzed the characteristics of defaulted bonds and obtained the association rules with the highest support, confidence, and lift values, as well as the attribute frequency ranking, between matured bonds and defaulted bonds. The association rules for risky bonds are shown in Table 11, and those for risk-free bonds are presented in Table 12. Based on the risk bond association rule, we observe that the credit risk of corporate bonds is more concentrated in four key industries: real estate, materials, discretionary consumption, and industrial sectors (Table 13). Utilizing association rules, we conduct a cross-statistical analysis of company attributes and issuer locations across these four industries. Overall, non-listed issuers exhibit higher risk levels compared to their listed counterparts in each industry.

The results of the cross-tabulation analysis of risk bonds with listing status and geographic distribution attributes for the industrial, real estate, consumer discretionary and materials sectors are shown in Figure 6. The risk of listed companies in the optional consumption sector is lower than that of unlisted enterprises, and unlisted private enterprises in South China and central state-owned enterprises in North China are particularly worthy of attention. The risks of corporate bonds in the real estate industry are primarily concentrated within private enterprises. Whether the issuing entity is listed or not has a negligible impact on risk discrimination. The bonds issued by non-listed private enterprises in North China, East China, and South China, as well as those issued by listed private enterprises in North China, South China, and Southwest China, entail relatively significant risks. Nevertheless, the bonds issued by non-listed Sino-foreign joint venture real estate enterprises have lower risks than those issued by wholly foreign-owned enterprises in the real estate sector. The bonds issued by non-listed Sino-foreign joint venture real estate enterprises in Northeast China encounter the least risk exposure.

The risk associated with bonds issued by the industrial and materials sectors is predominantly concentrated in unlisted enterprises. Within the industrial sector, unlisted private enterprises in East China, North China, and South China exhibit higher risk levels, bond defaults among centrally-owned state enterprises in North China warrant attention. In contrast, bonds issued by locally-owned state enterprises and foreign joint ventures tend to be less risky. The risk profile of bonds issued by industrial enterprises demonstrates a strong regional clustering effect. In the material industry, the number of bond defaults of private enterprises in East China is the highest, because the material industry is greatly affected by the energy policy, which leads to the cash flow turnover problem of bond issuers in the material industry, especially private material enterprises. The corporate bond risk shows no substantial variation based on regional or other characteristics.

The above analysis has examined the distribution characteristics of defaulted bonds from an industry perspective. It can be concluded that industries with a higher correlation to bond defaults mainly have strong cyclical nature, high leverage, policy sensitivity, and external environmental uncertainty. During economic downturns or deteriorating market conditions, the profitability, cash flow, and financing capabilities of these industries are more vulnerable to

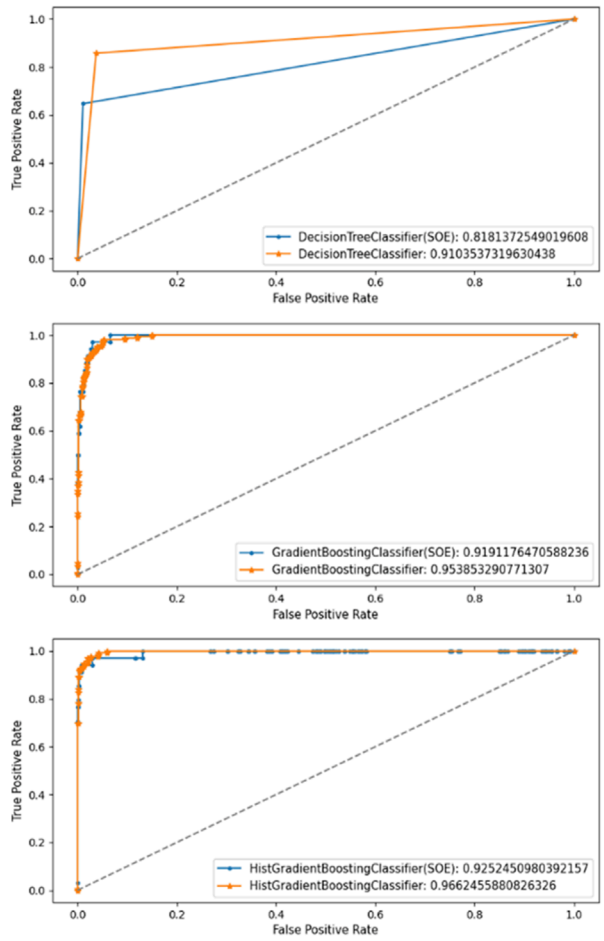

FIGURE 7
Comparison of ROC curves.

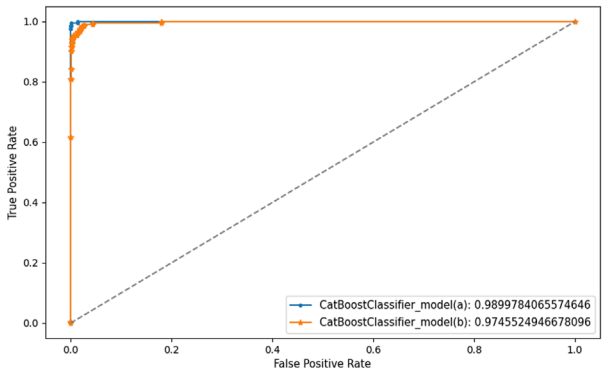

FIGURE 8
ROC curve for the ablation experiment.

shocks, thereby increasing the risk of default. From the perspective of whether they are listed or not, the average risk of bond issuance by listed companies is lower than that of unlisted companies. In terms of regional distribution, the characteristics of default regions are mostly concentrated in the eastern, northern, and southern parts of China, mainly due to the greater economic vitality and a larger number of enterprises in these regions.

The results of the above descriptive statistics are basically consistent with the analysis and association rule mining results, it confirms the effectiveness of association rule mining. The cross-analysis of the four industries can grasp the distribution and concentration of bond credit risk from a macro perspective, and assist investors to make relatively safe decisions when investing in public offering corporate bonds.

4.3 Bond default forecast

The attribute of enterprise frequently appears in the mining of association rules, and the results of association rules show that the bond risks of different industries vary in this dimension. Using machine learning models, we segmented the dataset into two groups based on the nature of the bond issuers: state-owned enterprises (SOEs) and non-state-owned enterprises (NSOEs). Bond risk prediction was then conducted on these two datasets, yielding distinct prediction outcomes. The results are shown in Table 14.

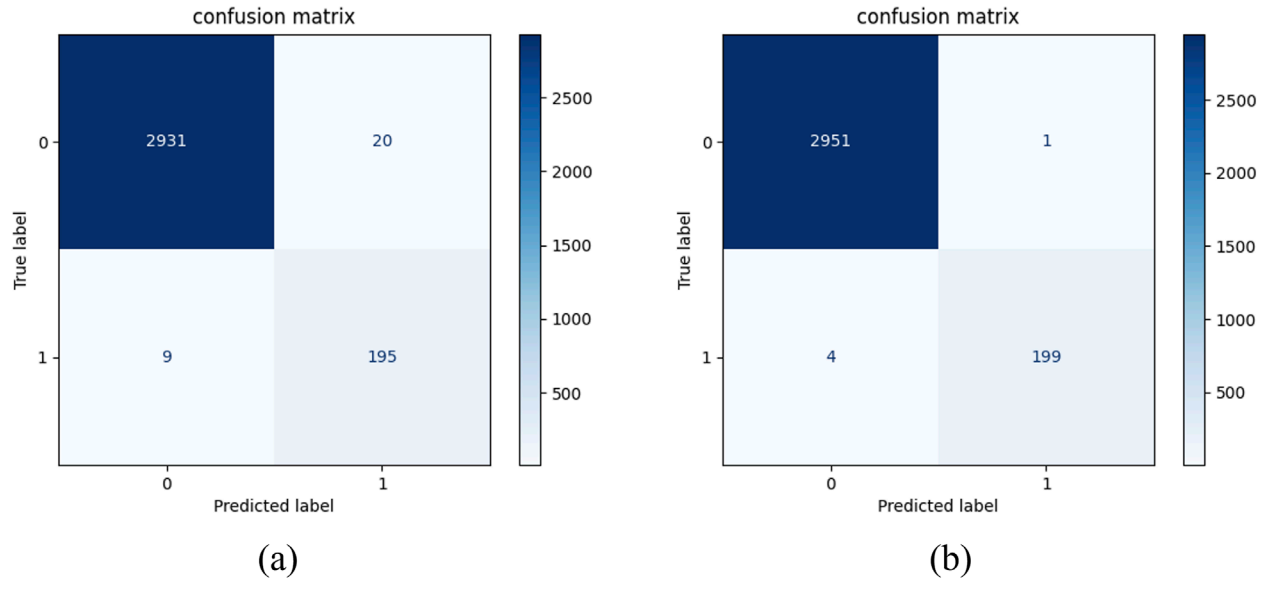

FIGURE 9
Comparison chart of confusion matrices for the ablation experiment. **(a)** Confusion matrix of Catboost before adding macro features; **(b)** Confusion matrix of Catboost after adding macro features.

By comparing the experimental outcomes of the two groups of data, it is observed that the accuracies of decision trees and GBDT in the two groups of experiments are not high. From the perspectives of precision, recall rate, and F1-score metrics, the prediction accuracies of random forests, HistGradientBoostingClassifier, and Catboost for non-state-owned bonds are higher than those for state-owned bonds. CatBoost performs best on the state bond dataset, while HistGradientBoostingClassifier and CatBoost are the top models for another group. According to the ROC curve analysis (Figure 7), the AUC values predicted by the decision tree and GBDT algorithm models for state-owned bonds are higher than those for non-state-owned enterprise bonds. Conversely, other models predict higher AUC values for non-state-owned bonds compared to state-owned enterprise bonds. Among the models, Random Forest and CatBoost exhibit similar performance across both datasets. The HistGradientBoostingClassifier achieves an AUC score that is 0.0916 higher on the non-state bond dataset compared to the state bond dataset. CatBoost performs best on the state bond dataset, while HistGradientBoostingClassifier outperforms other models on the non-state bond dataset.

In addition, to verify the conclusion of the var experiment, we selected the Catboost model and conducted ablation experiments on the indicators M1 and M2. The experimental results are shown in Figures 8, 9. The addition of the monetary indicator increases the AUC area of Catboost and reduces the number of bond misclassifications. This proves that m1, m2 has an enhancing effect on recognizing bond risk.

5 Results and discussion

The three experiments, VAR analysis, association rule mining, and descriptive cross-statistical analysis, analyzed the risks of credit bonds from the overall perspective of the bond market. Granger causality tests were conducted on the monetary indicators M1 and M2 and bond defaults, and the results confirmed the existence of a causal relationship between macroeconomic indicators and bond defaults. M1 and M2 have an impact on bond defaults: In light of the actual situation of China's bond market, China's interest rate liberalization is nearly complete, and the bond market has become an important platform for monetary policy operations. The interest rate transmission from the money market to the bond market is an important channel for monetary policy. M1 and M2 influence the bond market through liquidity effects and economic expectations, and their changes have an impact on the trend of the bond market. Bond defaults have an impact on M1: Against the backdrop of a high incidence of credit default events, defaulted bonds will lose liquidity due to trading suspension. Investors' risk aversion will lead them to choose to sell liquid bonds in their asset portfolios, thereby weakening the liquidity of high-quality bonds. Bond market risks will also lead to a decline in systemic risk preferences in the financial market, further affecting monetary policy.

Association rule mining and cross-analysis identified the current existence and concentration points of risks. Based on the results of association rule mining, we have identified four industries with concentrated bond risks: real estate, materials, consumer discretionary, and industrial. These four industries with high-risk correlation are highly consistent with the risk concentration points in China's bond market. The industrial and materials sectors have been affected by policies such as China's industrial structure upgrade and environmental protection production restrictions. As a result, some enterprises with overcapacity and high energy consumption and pollution have faced increased operational pressure, rising financing difficulties, greater capital turnover pressure, and continuous decline in profit levels, leading to accelerated bond risk release. The consumer discretionary sector is highly sensitive to economic cycles and is

TABLE 15 Feature importance ranking of catboost.

Feature	Importance ranking of state-owned bonds	Feature	Ranking of non-state-owned bonds
Net cash flows from operating activities/Non-current liabilities	33.40934	Net cash flows from operating activities/Non-current liabilities	40.15558
Net assets (year-on-year growth rate)	8.340914	Impairment loss on assets/Total operating revenue	5.10537
pca2	4.570496	Net non-operating income and expenses/Total profit	3.173927
Financial expenses/Total operating revenue	3.597293	Financial expenses/Total operating revenue	2.932994
Proportion of shares held by major shareholders	3.51573	Long-term asset suitability rate	2.766285
pca4	3.420728	Net assets (year-on-year growth rate) M1	2.420678
Net cash flows from operating activities as a percentage	3.065502	M1	2.365089
Asset impairment loss/Total operating revenue	2.900967	Total assets (year-on-year growth rate)	2.346182
pca5	2.535832	GDP	2.307728
Inventory turnover rate	2.440826	EPU	1.959694
pca1	2.311772	Current liabilities/Total liabilities	1.914631
GDP	2.085854	M2	1.774024
Current assets/Total assets pca7	1.986099	The shareholding ratio of major shareholders	1.758257
pca7	1.985943	pca2	1.746224
Accounts payable turnover ratio	1.789056	Local government debt	1.605419
Net cash flows from investing activities as a percentage of total cash flows	1.551838	Accounts payable turnover ratio	1.604
Net cash flows from financing activities as a percentage of total cash flows	1.477189	Tangible assets/interest-bearing debt	1.53493
Cash turnover ratio	1.398997	Proportion of net cash flows generated from operating activities	1.51413
Current liabilities/Total liabilities	1.364536	Proportion of net cash flows generated from investing activities	1.367484
Cash received from selling goods and providing services/Operating income	1.138044	Current assets/total assets	1.346553

mainly affected by the sluggish macroeconomy and weak consumer demand. For the real estate sector, it is because China's real estate policy regulation has become stricter in recent years. Regulations such as "houses are for living in, not for speculation" have suppressed housing price increases while also restricting the financing of real estate enterprises, leading to accelerated exposure of bond credit risks. According to cross-analysis, the risks of corporate bonds in the real estate industry are mainly concentrated in private enterprises, which confirms the current situation of two-tiered differentiation in the operation of Chinese real estate enterprises. Currently, the net bond financing of state-owned real estate enterprises in China has significantly increased, but the financing situation of some private real estate enterprises is not optimistic. At the same time, we have found that in each industry, the risk level of non-listed issuers is higher than that of listed companies. This is because listed companies have diverse financing channels, are subject to more supervision, and can

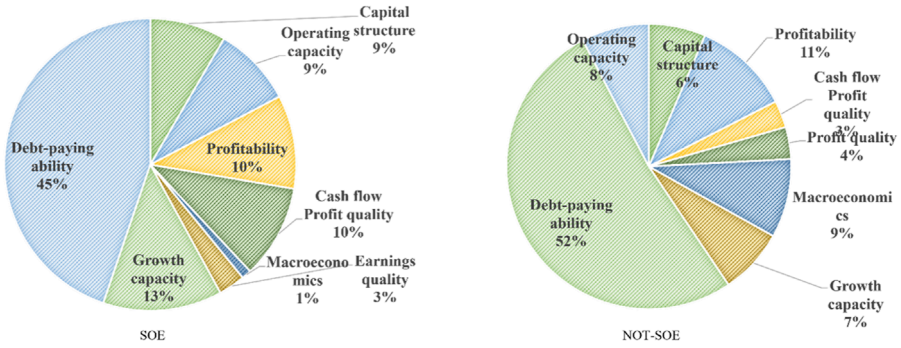

FIGURE 10
Comparison of risk factors.

disclose more information. Therefore, the credit bond risks of listed enterprises are usually smaller than those of non-listed enterprises. From a regional perspective, the economic development levels in South China and East China are relatively high, with a large number of private enterprises and concentrated bond issuers. Most enterprises are mainly engaged in trade, manufacturing, and export processing, and are more vulnerable to economic environment changes, thus leading to relatively concentrated default risks. These discussions provide valuable insights for subsequent focused attention and targeted adjustments in industries and regions experiencing frequent defaults.

With individual credit bonds as the research object and based on the aforementioned results of machine learning, we explore the influencing factors of state-owned and non-state-owned bonds, and further compare them for difference analysis. Since the AUC of Catboost is greater than 90% on both sets, we give the ranking of the top 20 feature importance of Catboost after removing the principal component index on the two data sets (Table 15). Through the feature importance weights, a comparison chart of risk factor weights for the two groups of experiments was obtained (Figure 10). Debt-paying ability can directly reflect the risk of credit bonds, and its proportion in private enterprises exceeds 50%. This indicates that financial indicators related to debt-paying ability most directly reflect bond risk. Capital structure also holds certain significance for the risk identification of credit bonds. This indicates that enterprises should pay attention to the investment and financing structure as well as the proportion of enterprise expenditures during the investment and financing process. A reasonable investment and financing ratio can maintain the stability of cash flow to diversify risks. For listed companies, the risk of bonds is smaller than that of issuing stocks, and it also provides them with a tool to diversify risks. Profitability reflects short-term solvency, while development capacity is demonstrated through indicators such as revenue growth rate, asset growth rate, and ROE, which also reflect long-term solvency. The combination of the two comprehensively reflects the solvency of an enterprise and represents its development potential, making them relatively important risk identification indicators. The proportion of macroeconomics in the two sets of data varies greatly. Non-private enterprises are greatly affected by macroeconomics and macro factors on their own operations. Private enterprises are more susceptible to policy adjustments. When the credit environment is relatively tight and the overall financing environment tightens, the liquidity pressure on private enterprises' funds will increase, which will significantly increase the risk of bond default. Therefore, issuers should make accurate predictions of the macro environment and the development positioning of their enterprises before issuing credit bonds.

6 Conclusion

This paper analyses bond default from the perspective of macro and micro combination. Based on the macro perspective, the logical transmission between macro factors and bond default is confirmed by causation, and the overall characteristics of bond default are grasped by the statistical research method combining association rule mining and descriptive statistics. Bonds are divided into no-risk bond group and risk bond group, and association rules are mined in four dimensions, corporate bond region, whether the issuing entity is listed, the issuing enterprise attribute, and whether there is a guarantee for corporate bonds, and rules with high correlation degree are obtained, the cross-analysis of risk factors was carried out in these four different dimensions. Microscopically, the bonds issued by enterprises were taken as the research object. The characteristics of defaulted bonds were analysed from the perspective of whether the issuer is a state-owned enterprise or not, and the key factors for bond risk identification and the integrated algorithm with the highest AUC were obtained. The conclusions are as follows:

(1) Through VAR and Granger causality test, we prove that bond default is correlated with M1 and M2;
(2) In view of the text attributes of bonds and bond issuers, correlation mining is conducted to find out the risk attributes strongly related to bond, and further conducted descriptive statistics on risk attributes to further grasp the risk characteristics and trends on a macro level.
(3) Bonds are grouped based on the risk correlation features obtained from association rules mining, and the credit status of bonds is predicted by combinational learning algorithm. The importance of risk attributes is ranked from the micro level according to the prediction results, and the validity of Granger causality test is further confirmed. At the same time,the combination learning model used in this paper has a high accuracy for bond default prediction, the accuracy rate of all

algorithms exceeds 90%. In terms of recall rate, the ensemble learning algorithm is better than the decision tree. The AUC of Catboost on both sample sets is greater than 90%. Therefore, we chose CatBoost to conduct a comparative experiment on the risk identification of bonds issued by state-owned and non-state-owned enterprises. By analysing the importance features of two groups of machine learning, the ranking of feature importance was obtained. The top-ranked feature was the solvency of enterprises, including important indicators such as "net cash flow from operating activities/non-current liabilities", indicating that the cash flow status of enterprises is extremely crucial for assessing the risk of corporate bonds. Macro factors are also key factors in bond risk identification. Non-state-owned enterprises are more affected by macro and their own operating conditions. In addition, in the ranking of the importance of default factors for non-state-owned bonds, both M1 and M2 are relatively important features. This results and the ablation experiments in machine learning indicate the correlation between bond default and M1 and M2, verifying the correctness of the aforementioned causal test from a micro perspective. Our research can provide a tool reference and guidance for risk management in the actual bond market.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

YWZ: Data curation, Formal Analysis, Investigation, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing–original draft, Writing–review and editing. WC: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

References

1. Merton RC. On the pricing of corporate debt: the risk structure of interest rates. *The J Finance* (1974) 29(2):449–70. doi:10.1111/j.1540-6261.1974.tb03058.x

2. Kedia S, Rajgopal S, Zhou X. Did going public impair Moody' s credit ratings? *J Financial Econ* (2014) 114(2):293–315. doi:10.1016/j.jfineco.2014.07.005

3. Yang B, Bin K, Li B, Julia Yu Y, Zhang J. Detecting accounting fraud in publicly traded US firms using a machine learning approach. *J Account Res* (2020) 58(1):199–235. doi:10.1111/1475-679X.12292

4. Beaver WH. Financial ratios as predictors of failure. *J Account Res* (1966) 4(12):71–111. doi:10.2307/2490171

5. Altman E, Fargher N, Kalotay E. A simple empirical model of equity- implied probabilities of default. *J Fixed Income* (2011) 20(3):71–85. doi:10.3905/jfi.2011.20.3.071

6. Douglas A, Huang AG, Vetzal KR. Cash flow volatility and corporate bond yield spreads. *Social Sci Electron Publishing* (2016) 46(2):417–58. doi:10.1007/s11156-014-0474-0

7. Sun J, Hui L, Fujita H, Fu B, Ai W. Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting. *Inf Fusion* (2020)(54) 128–44. doi:10.1016/j.inffus.2019.07.006

8. Beaver WH, Mcnichols MF, Rhie JW. Have financial statements become less informative? Evidence from the ability of financial ratios to predict bankruptcy. *Rev Account Stud* (2005) 10(1):93–122. doi:10.1007/s11142-004-6341-9

9. Sadiq M, Alajlani S, Hussain MS, Ahmad R, Bashir F, Chupradit S. Impact of credit, liquidity, and systematic risk on financial structure: comparative investigation from sustainable production. *Environ Sci Pollut Res* (2022) 29(14):20963–75. doi:10.1007/s11356-021-17276-x

10. Chava S, Jarrow RA. Bankruptcy prediction with industry effects. *Eur Finance Rev* (2004) 8(4):537–69. doi:10.1007/s10679-004-6279-6

11. Yuhui WU, Xinxin LIU, Chen Y. Optimization and improvement of bond default early warning model - based on SMOTE-tomek-GWO-XGBoost method. *Friends Account* (2024)(06) 73–81. doi:10.3969/j.issn.1004-5937.2024.06.011

12. Fama EF, Jensen MC. Separation of ownership and control. *J L Econ* (1983) 26(2):301–25. doi:10.1086/467037

13. Wilson BA. Movements of wages over the business cycle: an intra-firm view. *Finance and Econ Discussion* (1997) 1997.0:1–38. doi:10.17016/feds.1997.01

14. Jun LU, Zifang LI. Empirical analysis of the effect of macro economic factors to corporate financial crisis. *J Shanxi Univ Finance Econ* (2008)(11) 94–100. doi:10.3969/j.issn.1007-9556.2008.11.017

15. Chen H. Macroeconomic conditions and the puzzles of credit spreads and capital structures. *The Natl Bur Econ Res* (2010) 65(6):2171–212. doi:10.1111/j.1540-6261.2010.01613.x

16. Xiao LIU, Rongxi ZHOU, Yuru LI. Default prediction of credit bond in China based on stacking algorithm integrated model. *Operations Res And Management Sci* (2023) 32(03):163–70. doi:10.12005/orms.2023.0096

17. Fan LZ, Zhang C. Explanation of Macro economic variables on bond risk premia in China. *J Management Sci China* (2009) 12(06):116–24. doi:10.3321/j.issn:1007-9807.2009.06.013

18. Dai guoqiang, Sun xinbao. On the macro determinants of credit spreads of corporate bonds in China. *J Finance Econ* (2011) 37(12):61–71.

19. Zhang K, Liang Z, Liu Z, Jia N. A deep learning based multitask model for network-wide traffic speed prediction. *Neurocomputing* (2020) 396:438–50. doi:10.1016/j.neucom.2018.10.097

20. Luo H, Cai J, Zhang K, Xie R, Liang Z. A multi-task deep learning model for short-term taxi demand forecasting considering spatiotemporal dependences. *J Traffic Transportation Eng (English Edition)* (2021) 8(1):83–94. doi:10.1016/j.jtte.2019.07.002

21. Yunjie W, Shaolong S, Kin Keung L, Ghulam A. A KELM-Based 587 ensemble learning approach for exchange rate forecasting. *J Syst Sci Infor* (2018) 6(4): 289–301. doi:10.21078/JSSI-2018-289-13

22. Xu F, Keselj V. Collective sentiment mining of microblogs in 24-hour stock price movement prediction. In: *16th IEEE conference on business informatics (CBI)* (2014). p. 60–7. doi:10.1109/CBI.2014.37

23. Almalaq B, Dai X, Zhang J, Hanrahan S, Nedrud J, Hebb A. Causality graph learning on cortical information flow in Parkinson's disease patients during behaviour tests. *49th Asilomar Conf Signals, Syst Comput* (2015) 925–9. doi:10.1109/ACSSC.2015.7421273

24. Yanwei H, Dengguo WU, Jun LI. Structural healthy monitoring data recovery based on extreme learning machine. *Computer Eng* (2011) 37(16):241–3. doi:10.3969/j.issn.1000-3428.2011.16.082

25. Vazquez-Patino A, Samaniego E, Campozano L, Avilés A. Effectiveness of causality-based predictor selection for statistical downscaling: a case study of rainfall in an Ecuadorian Andes basin. *Theor Appl Climatology* (2022) 150(3-4):987–1013. doi:10.1007/s00704-022-04205-2

26. Li M, Zhang R, Liu K. Machine learning incorporated with causal analysis for short-term prediction of sea ice. *Front Mar Sci* (2021) 8:8. doi:10.3389/fmars.2021.649378

27. Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. *ACM SIGMOD Rec* (1993) 22(2):207–16. doi:10.1145/170036.170072

28. Schapire RE. The strength of weak learnability. *Machine Learn* (1990) 5(2):197–227. doi:10.1007/bf00116037

29. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. *Machine Learn* (2017). doi:10.48550/arXiv.1706.09516