Financial Innovation

RESEARCH

Open Access

An innovative machine learning workflow to research China's systemic financial crisis with SHAP value and Shapley regression

Da Wang[1] and YingXue Zhou[1*]

*Correspondence:
zhouyingxue1996@163.com

[1] School of Economics, Jilin University, 2699 Qianjin Street, Changchun, Jilin Province, China

Abstract

This study proposed a cutting-edge, multistep workflow and upgraded it by addressing its flaw of not considering how to determine the index system objectively. It then used the updated workflow to identify the probability of China's systemic financial crisis and analyzed the impact of macroeconomic indicators on the crisis. The final workflow comprises four steps: selecting rational indicators, modeling using supervised learning, decomposing the model's internal function, and conducting the non-linear, non-parametric statistical inference, with advantages of objective index selection, accurate prediction, and high model transparency. In addition, since China's international influence is progressively increasing, and the report of the 19th National Congress of the Communist Party of China has demonstrated that China is facing severe risk control challenges and stressed that the government should ensure that no systemic risks would emerge, this study selected China's systemic financial crisis as an example. Specifically, one global trade factor and 11 country-level macroeconomic indicators were selected to conduct the machine learning models. The prediction models captured six risk-rising periods in China's financial system from 1990 to 2020, which is consistent with reality. The interpretation techniques show the non-linearities of risk drivers, expressed as threshold and interval effects. Furthermore, Shapley regression validates the alignment of the indicators. The final workflow is suitable for categorical and regression analyses in several areas. These methods can also be used independently or in combination, depending on the research requirements. Researchers can switch to other suitable shallow machine learning models or deep neural networks for modeling. The results regarding crises could provide specific references for bank regulators and policymakers to develop critical measures to maintain macroeconomic and financial stability.

Keywords: China, Machine learning, SHAP value, Shapley regression, Systemic financial crisis

Introduction

The coronavirus disease (COVID-19) pandemic posed a significant threat to the global economy. Therefore, preventing financial collapse based on new technical methods and comprehensive case studies is of great significance in the post-pandemic period, when

an uneven global rebound is more likely to challenge the financial stability of emerging markets and developing economies (Nivorozhkin and Chondrogiannis 2020). Moreover, prevention and monitoring of systemic financial risk in China is a top priority.

First, as the second-largest market in the world, China has become deeply integrated into the global economy and is now one of the critical driving forces of global economic growth. Consequently, attention focused on the Chinese financial system has increased globally. Second, the report of the 19th National Congress of the Communist Party of China demonstrated that China was confronted with severe challenges. It stressed that the government would strengthen monitoring, early warning, mitigation, and control of financial risks to ensure that no systemic risks would emerge. Thus, it is essential to monitor China's systemic risks to both the country and other economies.

Some empirical studies on systemic risk have used conventional economic models, including regression, but the performance of the predictions has not been excellent (Nag and Mitra 1999; Franck and Schmied 2003; Roy 2009; Sekmen and Kurkcu 2014; Tölö 2020). Therefore, some academics have proposed machine learning models to improve the prediction of systemic risk (Joy et al. 2017; Carmona et al. 2019; Tölö 2020). However, new problems have emerged because most of these models are non-parametric, non-linear and non-human-readable. Some cutting-edge research has employed additional explanatory models (Son et al. 2019; Suss and Treitel 2019; Bluwstein et al. 2020) following machine learning models to describe the model's results; however, the decomposition is only a part of the explanation of the model. We also need statistical inference as a hypothesis test to evaluate the confidence in a specific model's output (Joseph 2019). To the best of our knowledge, this is unusual in China's systematic risk research; therefore, we applied a new workflow for this purpose.

The workflow comprises three steps: relative model evaluation, decomposition of projected values into feature contributions, and statistical inference of feature attributes. First, modeling was conducted using machine learning techniques. Unlike most theoretical and econometric models, machine learning modeling does not require preset causality or specific constraints. However, they can provide higher accuracy and enable high-dimensional data processing, which is a shortcut to traditional methods. However, conventional methods are easier to interpret than machine learning models. Because machine learning models are highly complex and non-parametric, it is difficult to extract decision rules from them. There are, however, additional interpretation models for machine learning that increase the model's transparency. Compared with traditional models, machine learning models can learn more information and make the best use of data (Tölö 2020), which gives the interpretive model the opportunity to explain more to users.

Second, in this workflow, SHAP (SHapley Additive exPlanations) was selected to explain the findings of machine learning models. SHAP is a powerful approach that was developed to explain the output of any machine learning algorithm at the global and local levels. Locally, they explain why a given observation is assigned to a class, the contribution of each variable, and whether the effect is positive or negative. Globally, they estimate each variable's overall contribution and direction toward the target (Ariza-Garzón et al. 2020). Compared with the feature importance interface in Python, which is only used for the total feature importance ranking in training, SHAP

is extractable, additive, and comparable, with excellent flexibility. Furthermore, the direction and size of the SHAP value can be visualized, improving the intuitiveness and readability of the interpretation (Lundberg et al. 2018).

Third, statistical inference was conducted using Shapley regression. This enabled us to validate the feature-to-label alignment in the trained model by establishing local and linear regression procedures in the additive parameter space. The SHAP value approach creates space by transforming non-linear and non-extractable prediction functions constructed using machine learning models. It extends the parametric statistical inference to non-linear, non-parametric models. This is the only general framework for jointly conducting rigorous statistical inferences on all parameters in these complex models (Buckmann et al. 2021), which could expand the application of machine learning (Suss and Treitel 2019; Bluwstein et al. 2020; Joseph et al. 2021; Buckmann and Joseph 2022).

The workflow covers the parts from modeling to interpretation to statistical inference, with high predictive performance and explanatory transparency advantages. However, there is another crucial step in machine learning pipelines: feature engineering. Feature engineering uses domain knowledge to extract features from raw data to better fit the input data to the model (Zaidi 2015) and is frequently used for the transformation of unstructured data and dimensionality reduction of high-dimensional data. It is essential for the success of machine learning (Locklin 2014), and the correct features can ease the challenge of modeling, enabling the pipeline to output results of higher quality (Zheng and Casari 2018). We also note that most indicators chosen in the literature on China's systemic risk studies are based on theoretical analyses and prior knowledge, which are not conducive to the performance of machine learning models. Therefore, we updated this workflow by integrating the feature engineering phase before modeling rather than just the predetermined phase. This is more objective, yields more accurate results, and may yield unforeseeable indicators.

To select the indicators, we added a global trade factor. International credit is typically chosen in existing literature. However, Cesa-Bianchi et al. (2019) showed that the global trade factor has a modest effect on the domestic risk of countries that are more open to trade. Considering that China is open to trade, we expected foreign risks to affect China's domestic risks through international trade; therefore, we added global trade factors to represent the international environment. We then asked the model to determine whether this factor was useful for the prediction.

The training dataset of this study was obtained from the Jordà–Schularick–Taylor Macrohistory Database (Jordà et al. 2017, 2019) from 1870 to 2016, one of the most extended macroeconomic databases available. More importantly, it summarizes each country's systemic financial risk over a long period and provides additional risk characteristics for our study. Although it includes 17 countries whose production accounts for more than 90% of the output of developed countries and more than 50% of all countries worldwide, it does not only include these countries as a representative sample. Simultaneously, the indicators come from various sources, including many macroeconomic characteristics such as GDP, money, and interest rates, as well as financial parameters containing bank credit and returns on all types of assets, giving us the opportunity for feature engineering. Several meaningful studies have been conducted using this database. For example,

Schularick and Taylor (2012) and Bianchi (2020) used this database to make significant progress in research on risk and depression.

This study aimed to upgrade a new workflow and combine it with macro data to depict the probability of China's systemic financial crisis (the definition is given in "Datasets" section) and analyze the macro-level risk drivers and their confidence. Specifically, the main process is as follows: We selected the feature subset with the best performance of the risk model with feature engineering and ultra-long-term risk data from 16 countries. On this basis, we determined the optimal model parameters by cross-validation (CV) and then used this model to calculate the risk probability of China from 1990 to 2020. Finally, we decomposed the probability into the contribution of each feature by the interpretation model and calculated its confidence via statistical inference.

In summary, this study contributes to the literature in three ways. First, with respect to methodology, we upgraded a new workflow, which now involves four steps: objective selection of indicators, relative model assessment, decomposition of predicted values into feature contributions, and statistical inference of feature attributions. Compared with the traditional workflow, combining these four moves is more conducive to recognizing significant features and enhancing the quality of the subsequent model. Moreover, this routine serves as a firm reference for other researchers. It can be applied to any prediction and regression problem, regardless of the field. Second, because feature selection and non-parametric, non-linear model statistical inference are rare in systematic risk studies in China, much less a combination of the two, this study is an essential attempt to address this topic. Third, regarding the data and results, we considered an ultralong sample period, and the risk information was more comprehensive and richer, giving us the opportunity to capture more risk features and select more representative indicators. Besides, we added an international trade factor for feature selection. The findings reveal that this factor is conducive to measuring China's risk. Furthermore, we elucidated and quantified the possible non-linear and non-parametric relationships between systemic financial risk and country-level macroeconomic factors. Finally, we determined a safe haven zone that may benefit policymakers.

The balance of this paper is structured as follows: The second section, "Literature review," summarizes relevant literature from the perspectives of methods and indicators. The third section, "Datasets" describes the targets, variables, and samples used in this study. Section "Workflow and methodology" introduces the upgraded workflow and methods selected in this article. We proceed in the following order of workflow steps: "Feature engineering", "Building the model", "Decomposing model results", and "Statistical inference on the decomposition" sections. The fifth part, "Workflow performance" section shows the empirical results in the global dataset and the sixth part, "Trained workflow in China's risk" presents in-depth details of its risks. In addition, the results presented in the fifth and sixth sections are organized in the workflow order. Finally, Section "Discussion and conclusion," concludes the paper.

Literature review

Methods

Many early warning models have been proposed to recognize and predict systemic risk after the outbreak of regional financial crises in the 1990s. For example, the probit model

(Frankel and Rose 1996), Kaminsky–Lizondo–Reinhart (KLR) signal analysis (Kaminsky et al. 1997; Kaminsky 1999; Shi and Gao 2009), and the cross-section regression Model (Sachs et al. 1996) have been extensively used. Nevertheless, the cross-section regression model cannot compute the accurate risk probability, and the KLR signal analysis is subjective in setting indicators and thresholds.

Consequently, more models have been studied, beginning with the developing country studies division (DCSD) model (Berg and Pattillo 1999) combining probit with KLR, followed by the logit model (Su and Xiao 2011) and synthetic index technique (Illing and Liu 2006; Cardarelli et al. 2011; Tao and Zhu 2016; Tsinghua University Research Team 2019). Additionally, the conditional value-at-risk (CoVaR) (Chen and Wang 2014; Adrian and Brunnermeier 2016; Fang et al. 2018), Expected Shortfall (Fan et al. 2011; Acharya et al. 2017; Yang et al. 2018, 2019), and CoRisk (Chan-Lau et al. 2009) have been widely applied. Other innovative techniques are based on assets and liabilities (Greenwood et al. 2015; Fang 2016).

Bisias et al. (2012) divided systemic risk measurement techniques into six groups: macroeconomic measures, granular foundations, network measures, forward-looking risk measurements, stress tests, cross-sectional measures, and measures of illiquidity and insolvency, including 31 specific methods that are systematically reviewed and compared. Billio et al. (2017) also introduced systemic risk measurement models. Generally, traditional risk identification models have considerable advantages in interpretation but are restricted to the distribution and dimensions of the input data and may show poor predictability.

Therefore, academics have introduced machine learning models into risk analysis. Many scholars have conducted studies on supervised learning. For instance, Ecer (2013), Eygi (2013), Jones et al. (2015), and Ekinci and Erdal (2017) demonstrated the validity of supervised learning for risk identification. Some scholars, such as Li et al. (2022), provided research on risk using unsupervised learning. In particular, the performance of deep neural networks, which fall into the category of supervised learning, is much better; however, they require much data, limiting their application in the macroeconomy and finance. Nevertheless, the most advanced technologies in computer vision, such as One-Shot Learning, Meta-Learning, and Transfer Learning, aim to simulate human brains and let the model learn "learning," enabling us to conclude rapidly from limited data and enhance the model generalization ability, which increases the possibilities to apply deep neural networks in economics and finance and shows a direction for our future research. Kou et al. (2019) surveyed existing research and methodologies for the assessment and measurement of financial systemic risk combined with machine learning technologies and identified future challenges, as well as suggesting further research topics.

This study selected a prediction model from two widely used supervised learning models: random forest (RF) and gradient boosting decision tree (GBDT). First, RF has made considerable progress in risk prediction. For example, Ward (2017) used RF to predict systemic banking risk under two sample times, including from 1870 to 2011 and after 1970, with very good results; Suss and Treitel (2019) developed an early warning system of banking systemic risk with FSA's bank-scoring database and showed that the performance of RF was better than conventional models and other machine learning algorithms; Wang and Zhou (2020) also revealed that in systemic risk identification, machine

learning models are significantly better than conventional models both in terms of learning and risk recognition abilities, and machine learning models are more stable; Wang (2019) constructed an experimental study with data of Global Financial Crises from 1970 to 2011 and provided evidence that RF was effective at both identifying the leading indicators and risk prediction.

Second, researchers at the Bank of England and other major central banks gradually applied GBDT to macroeconomic and financial research. For instance, Momparler et al. (2016) applied this approach to predict the failure of commercial banks in 155 Eurozone countries between 2006 and 2012 and derived four useful indicators. Carmona et al. (2019) concentrated on a performance comparison and showed that GBDT outperformed logit and RF in 156 US commercial banks' bankruptcy predictions between 2001 and 2015. Zięba et al. (2016) highlighted that the GBDT showed the best performance in binary classification, such as crisis identification, compared with traditional linear and most machine learning models.

According to Altmann et al. (2020), additional interpretation models can be classified according to various criteria, including whether they are model-specific or model-agnostic, and whether they explain the individual prediction (local) or the entire model behavior (global). There are several popular methods belonging to Global Model-Agnostic Methods, such as the partial dependence plot (PDP; Friedman 2001) and permutation feature importance (Fisher et al. 2018). In addition, Local Model-Agnostic Methods include local interpretable model-agnostic explanations (Ribeiro et al. 2016), individual conditional expectations (Goldstein et al. 2015), and explanation vectors (Baehrens and Schroeter 2010). Johansson et al. (2004), Barbella et al. (2009), Florez-Lopez and Ramon-Jeronimo (2015), and Son et al. (2019) investigated the application of different interpretation methods to different machine learning algorithms with superior findings.

Based on cooperative game theory, the interaction-based method of explanation (IME) is more suitable for complex interactions between variables and targets (Strumbelj and Kononenko 2010). This is a clear characteristic of economic and financial research. The SHAP value (Lundberg and Lee 2017) is a typical IME model, and recent investigations have shown its advantages in interpreting complex models in economic studies (Chakraborty and Joseph 2019; Bracke et al. 2019; Wang et al. 2020) and in feature selection (Chu and Chan 2020). Simultaneously, numerous recent criticisms and concerns of the Shapley-value-based explanations in machine learning have indicated that we should have a better understanding of its limitations (Kumar et al. 2020; Chen et al. 2020).

Additionally, the Bank of England proposed the Shapley regression method (Joseph 2019) based on the SHAP value to further test the economic and statistical significance of the features. Joseph (2019) illustrated the feasibility of this framework in machine learning by proving the polynomial consistency of machine learning estimators and the composition bias theorem of Shapley values. Furthermore, Joseph highlighted two conditions that make the inference valid: independence between the model optimization and coefficient estimation and sufficiently fast convergence of the model. Both can be addressed through unbalanced sample splitting between the training and test sets, as is typical in machine learning applications.[1] It has been successfully used to predict

[1] Sample splitting during training process was achieved by cross-validation in this paper due to the small size of data.

long-term systemic risk (Bluwstein et al. 2020), early warning systems for bank distress (Suss and Treitel 2019), US unemployment (Buckmann and Joseph 2022), and UK inflation (Buckmann et al. 2021). This is the crucial theoretical basis of the present study.

Furthermore, Zheng and Casari (2018) stated that feature engineering is a crucial step in machine learning pipelines and provided a detailed description. Heaton (2016) demonstrated that most models performed differently for different engineering features. In economics, the least absolute shrinkage and selection operator (LASSO) is frequently used for feature selection. Yan et al. (2020) used LASSO to design an early warning system for financial distress experienced by listed companies. However, many more methods and ideas can be borrowed from machine learning. For instance, Heiberger (2018) applied the recursive feature elimination (RFE) method to project economic trends using listed company stock data, providing much help by yielding more compact subsets of features and eliminating information redundancy. Therefore, we selected the RFE method for this study. In addition, Kou et al. (2021) proposed a two-stage, multi-objective feature selection method to determine the final feature subsets, and the results showed that the proposed model achieved similar and sometimes better performance when using far fewer features than multi-objective, wrapper-based feature selection methods and other benchmarks. This coincides with the idea of this study, where we also adopted the approach of first filtering the features with a univariate filtering method and then combining the model to select the final subset among the remaining subsets.

Indicators

Since we predicted risk using macroeconomic data, we found a summary from the relevant literature that macropredictors of risk can be roughly divided into the following categories[2]:

(1) Economic performance: Economic health significantly impacts the financial market, and overheating or a recession can pose risks. Representative indicators include the GDP growth rate, inflation rate, and Consumer Price Index (Frankel and Saravelos 2012; Caggiano et al. 2014; Schularick and Taylor 2012).
(2) Fiscal: Fiscal indicators reflect government revenue or liabilities and are more likely to be at risk if the government spends more than it earns. The indicators usually used in the literature are the ratio of government spending to revenue, government revenue to GDP, and the ratio of government liabilities to income (Tao and Zhu 2016; Laeven and Valencia 2013).
(3) Money: Monetary indicators are directly connected to financial markets. For instance, deposit and loan interest rates influence the credit scale and broad money reflects market liquidity, both of which are highly sensitive to risk. In addition, broad money can also measure the degree of financialization of a country and whether there is a bubble economy, which is closely connected to risk (Kaminsky et al. 1997; Kaminsky 1999; Caggiano et al. 2014).
(4) Real estate: Asset price bubbles led by the real estate market are a characteristic feature of recent financial crises. For instance, the 2008 Global Financial Crisis was

[2] To be clear, here we focus on classifying the indicators used in the literature and analyzing the association of each category with risk. The link between the final indicators and risk can be seen in "Feature selection and evaluation" section.

triggered by a decline in the quality of real estate loans. Real estate prices, commonly represented in the literature by the one-year growth rate of real house prices, affect a financial system's stability through financial leverage. The higher or faster house prices rise, the more likely it is that risk will occur (Joy et al. 2017; Tölö 2020).

(5) Financial submarkets: Financial submarkets are locations for the direct transmission of risk, which is of great significance for systemic risk prediction. Loan-to-deposit ratios and nonperforming loans are often employed to measure banking system performance (Tao and Zhu 2016; Lainà et al. 2015). The stock market is measured using the stock price and overall market value of listed companies (Tao and Zhu 2016; Tölö 2020), interest rates on long-term or short-term government bonds, and the slope of the yield curve are representative of the bond market (Joy et al. 2017).

(6) Balance of payments: This group refers to how closely the domestic economy is connected to the global economy, including imports, exports, foreign assets, and the current account or its ratio to GDP, which can affect capital flow between home and abroad. The more closely a country is linked to the global economy, the more vulnerable it is to external risks. For example, a persistent current account deficit hinders the growth of the domestic economy, and long-term foreign debt affects a country's reputation and may exacerbate risks. In contrast, if the current account continues to be in surplus, the domestic economy becomes highly dependent on foreign capital, increasing economic instability and financial risks (Reinhart and Rogoff 2008; Su and Xiao 2011; Frankel and Saravelos 2012; Tölö 2020).

(7) External impact: This category represents the vulnerability of the world economy. Nations are now closely financially linked, and risk spillovers across borders are clear. Generally, the worse the performance of risk-sensitive factors abroad, the more likely the risk is in the country. In other words, the probabilities of threats at home and abroad are highly correlated. Presently, the literature has widely used the exchange rate, global GDP, global credit, and global slope of the yield curve as external impact factors (Kinkyo 2019; Cesa-Bianchi et al. 2019; Lo and Peltonen 2013; Bluwstein et al. 2020; Abbritti et al. 2018). However, the existing literature indicates that countries with open trade are not suitable for the global credit indicator without further extension. Considering that China is much more open in terms of trade than finance, global trade can be considered as a measure of the global environment.

The abovementioned database includes primary economic and financial data, allowing us to calculate other common indicators, and the obtained indicators can cover the above seven categories. Feature engineering was conducted to select the feature subset with the best prediction performance.

Overall, there is room for further investigation into the systemic risk. For example, most existing systemic risk literature has chosen variables based on theory and experience, which are subjective and may not lead to the best results. In addition, global trade factors have not been fully considered in countries that are open to trade. Besides,

regarding studies with machine learning techniques, because of the "black box" nature[3] of machine learning, most current risk research fails to detail the internal function of the model (i.e., how a variable impacts the dependent variable) or the statistical inference. These are the three aspects in which we are trying to make a breakthrough.

Datasets

In addition to the abovementioned database, we collected data on China from the World Bank and IFS databases.

Systemic financial crises are defined by the Jordà–Schularick–Taylor Macrohistory Database as "events during which a country's banking sector experiences bank-runs, sharp increases in default rates accompanied by large losses of capital that result in public intervention, bankruptcy, or forced merger of financial institutions" (Schularick and Taylor 2012). It is a binary label, named "crisisJST" in the database. The target of the training set determined the type of crisis captured by the model; thus, when applied to China, the model captured the same type of risk. Because the database used for training does not contain China's risks, we are concerned that the definition of risk in other databases may differ from that in the training dataset, which may cause errors in checking the validity of the model to identify China's risks. Therefore, instead of selecting Chinese risk data from other databases for numerical validation, we directly used Chinese indicators and trained models to calculate Chinese risk probabilities and validate the identification of risk probabilities based on relevant literature and the reality in China.

For the indicators, 22 original variables remained after data pre-processing (see Table 2). After feature engineering, 10 features (listed in Table 4) were selected.

For the training samples, based on the selection of the above database and countries, we performed some sample processing. Specifically, we set one and two years before the crisis as the positive samples (the lead-lag method[4]). Following Bluwstein et al. (2020), we eliminated the crisis year and the four years immediately following the crisis to prevent the interference of economic recovery on the risk prediction, which could also filter the noise for the model. In addition, we eliminated samples from the post-Great Depression economic recovery period of 1933–1939 and the two world wars of 1914–1918 and 1939–1945, as well as samples with missing values. Finally, 1005 samples were obtained, including 912 non-risk and 93 risk samples, and Table 1 summarizes the crisis years.

Workflow and methodology

This section explains the updated workflow and techniques employed at each stage, organizing it into a mind map, as shown in Fig. 1. The green square is the workflow step, and the white square is the method used in this study.

[3] The black-box nature of machine learning models refers to the fact that the model is highly complex, and we do not obtain the interactions between the different features within the model. As Molnar (2023) indicated, "black box" describes models that cannot be understood by looking at their parameters. Savage (2022) also pointed out that the decision-making process of a machine learning model is often referred to as a black box—researchers and users typically know the inputs and outputs, but it is hard to see what's going on inside.

[4] Generally speaking, the methods to study time series by machine learning include sliding window, lead-lag, and time series clustering, etc. Also there are other machine learning algorithms which itself has taken time into account such as recursive neural network and online learning (Chakraborty and Joseph 2019), but their computational complexity and hardware requirements are both very high due to the need of constantly updated data, so they are suitable for high-frequency and large-scale dataset. Considering data scale of this study and the data detrending in the latter, lead-lag is more suitable for us.

Table 1 Crisis years in samples

Country	Crisis years in samples
Australia	1989
Belgium	2008
Denmark	1885, 1908, 1931, 1987, 2008
Finland	1921, 1931, 1991
France	1930, 2008
Germany	1891, 1901, 1907, 1931, 2008
Italy	1990, 2008
Japan	1920, 1927, 1997
Netherlands	1907, 2008
Norway	1899, 1922, 1931, 1988
Portugal	2008
Spain	1977, 2008
Sweden	1878, 1907, 1922, 1931, 1991, 2008
Switzerland	1991, 2008
UK	1974, 1991, 2007
USA	1893, 1907, 1929, 1984, 2007

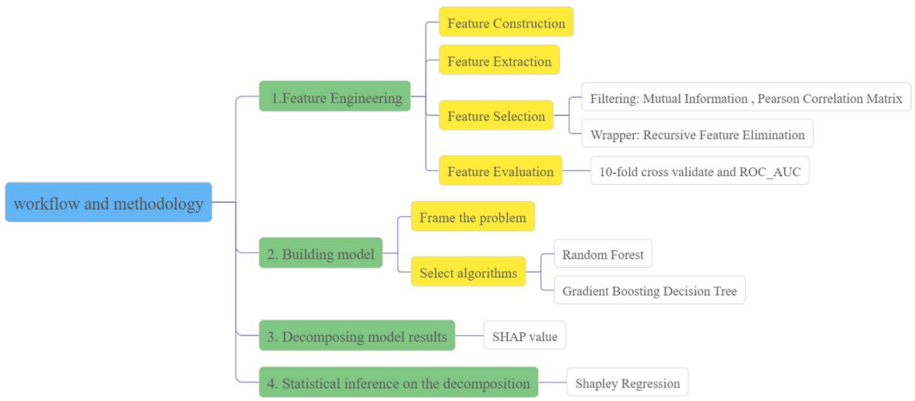


Fig. 1 Flowchart

Feature engineering

This process aims to determine the best subset of features from the original datasets to improve the model performance for classification or regression (Zheng and Casari 2018). It can be divided into four steps—feature construction, extraction, selection, and evaluation—which can be freely combined. In most cases, the dataset is large and chaotic; therefore, determining a subset of features usually requires an iterative spiral of the above stages. Alternatively, feature engineering is a highly iterative and repeated trial and error process.

Feature construction

Feature construction involves discovering missing information regarding the relationships between features and augmenting the space between them by inferring or creating

Table 2 Variables

Original variables	Type	Categories	Original variables	Type	Categories
rgdpmad	Level	Economic performance	eq_tr	Composite calculated	Financial submarkets
GDP	Level	Economic performance	housing_capgain	Composite calculated	Real estate
ca	Level	Balance of payments	eq_capgain	Composite calculated	Financial submarkets
imports	Level	Balance of payments	rgdppc	index	Economic performance
exports	Level	Balance of payments	rconpc	index	Economic performance
narrowm	Level	Monetary	cpi	index	Economic performance
money	Level	Monetary	iy	Ratio	Economic performance
revenue	Level	Fiscal	debtgdp	Ratio	Fiscal
expenditure	Level	Fiscal	eq_dp	Ratio	Financial submarkets
tloans	Level	Financial submarkets	eq_div_rtn	Ratio	Financial submarkets
stir	Level	Monetary	ltrate	Level	Monetary

more features. For numerical features, such as economics, researchers usually use a combination or decomposition of initial features by simple algebraic operators, such as average, addition, subtraction, multiplication, division, and their combinations (Motoda and Liu 2002).

The methodology employed in this paper at this stage:

(1) Calculate eight indicators using variables mentioned above while taking into account the knowledge of economics and with the economic significance considered: yield curve slope (difference between long-term and short-term interest rate), degree of foreign trade dependence (100 times of sum of import and export over GDP), the balance of trade (difference of export and import), commercial bank fixed deposit (difference of broad money and narrow money), bank loan-to-deposit ratio (ratio of credit to the commercial bank fixed deposit), cost of the long-term loan (total loans times long-term interest rate), the ratio of government spending to revenue and elasticity coefficient of government expenditure and revenue, shown in Table 3.

(2) We attempted to eliminate the time trend for all the above indicators based on the following principles[5]:

(i) Indicators developed using differencing, such as the balance of trade, calculate its ratio over GDP and the one-year difference in the ratio.

(ii) Index indicators, including CPI and GDP, calculate the one-year difference and annual growth rate of the index.

[5] All indicators and their corresponding processing methods were organized in Additional file 2: Appendix S1.

Table 3 Constructed variables

Constructed variables	Types	Categories
Degree of foreign trade dependence	Ratio	Balance of payments
Loan-to-deposit ratio	Ratio	Financial submarkets
Ratio of government spending to revenue	Ratio	Fiscal
Balance of trade	Level	Balance of payments
Commercial bank fixed deposit	Level	Financial submarkets
Cost of long-term toan	Level	Financial submarkets
Elasticity coefficient of government revenue	Composite calculated	Fiscal
Elasticity coefficient of government expenditure	Composite calculated	Fiscal
Yield curve slope	Level	Monetary
Global-minus	Composite calculated	External impact
Global-sum	Composite calculated	External impact
Global-im	Composite calculated	External impact
Global-ex	Composite calculated	External impact

(iii) The level indicators like imports, calculate its annual growth rate, a ratio over GDP, and the one-year difference in the ratio.[6,7]

(iv) For elastic coefficient index: no obvious trend, retain the original value.

(v) Ratio indicators and other composite-calculated indicators retain their original values and calculate the one-year differences.

(3) The global factor is calculated to represent the international environment. There are four global trade factors: Global-minus, Global-sum, Global-im, and Global-ex. The calculation formula for each country $c, c = 1, 2, \ldots N$ in year y is as follows:

$$GlobalA_{c,y} = \frac{1}{N-1} \sum_{i \in N, i \neq c} DomesticA_{i,y} \tag{1}$$

where $DomesticA_{i,y}$ represents the indicator of country i in year y, which we select as the net export to GDP ratio (Global-minus), the sum of imports and exports to GDP ratio (Global-sum), the import to GDP ratio (Global-im), and the export to GDP ratio (Global-ex).

Feature extraction

Feature extraction is a process that extracts a set of new features from original features via functional mapping (Motoda and Liu 2002). It transforms arbitrary unstructured data, including text or images, into numerical features that can be used for machine learning modeling (Zheng and Casari 2018). Furthermore, principal component analysis (PCA), independent component analysis, linear discriminant analysis, and other techniques are crucial for reducing the dimensionality (Khalid et al. 2014).

This study did not involve data such as text or images. Moreover, using methods such as PCA to synthesize indicators considering the following clearer interpretation is inappropriate; therefore, we did not work at this stage.

[6] Considering the economic relevance, the Yield Curve Slope only maintains the original value.

[7] We also calculated the one-year difference value for such indicators, but most indicators still showed an obvious time trend after this processing, so we abandoned this method.

Feature selection

Feature selection is a process that selects a subset of features from the original features such that the feature space is optimally reduced based on a particular criterion (Motoda and Liu 2002). Feature-selection techniques can be classified into three categories: filtering, embedded, and wrapper. Among them, the first evaluates only the goodness of a single variable, which is different from the latter two with the involvement of the learning algorithm, in which the wrapper technique performs the best. Typically, filtering first chooses, followed by embedded or wrapper methods for accurate feature selection (Zheng and Casari 2018).

The methodology applied at this stage in this study is mutual information (MI), the Spearman correlation matrix, which belongs to the filtering method, and RFE, which belongs to the wrapper class. An introduction to MI and RFE is provided below.

MI, which has good generalizability, captures the linear and non-linear correlations between features and labels. The MI score is calculated using the joint probability, which can be regarded as the probability of features and labels emerging simultaneously. Consequently, the correlation between the characteristics and labels gradually shifts from completely uncorrelated to perfectly correlated, as indicated by the value ranging from 0 to 1, which signifies that the two never come together, or always appear together. This study adopted the following calculation formula for discrete distributions considering the discreteness of the label[8]:

$$I(X,Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\,log\left(\frac{p(x,y)}{p_1(x)p_2(y)}\right) \tag{2}$$

where $p(x,y)$ represents the joint probability distribution function of X and Y, $p(x)$, and $p(y)$ are the marginal probability distribution functions of X and Y, respectively.

RFE (Guyon et al. 2002) is an extensively used feature selection technique in machine learning. It belongs to the wrapper category, which is the best way to enhance the model performance relative to filtering and embedded. This applies when the algorithm is determined because feature selection and model training are performed concurrently. It chooses features by recursively fitting the model and then removing the weakest characteristics (the number can be set manually) until the specified feature number (which also can be given in advance) is attained or the entire feature set is explored. The optimal feature subset was selected as that with the highest CV score (Wu et al. 2017). It is worth noting that RFE needs the feature importance ranking returned by the learner when removing features, so there are specific requirements for the training models.

[8] The calculation formula for continuous distributions:$I(X,Y) = \int_Y \int_X p(x,y) log\left(\frac{p(x,y)}{p(x)p(y)}\right) dxdy$

Feature evaluation

Feature evaluation is typically conducted along with feature selection to assess the performance of a currently chosen subset of features. The assessment criteria included the accuracy of classification and MSE for regression (Motoda and Liu 2002). After the experiments and comparisons, the negative log-likelihood loss[9] under the fivefold cross-validation was used, which was more stable on this dataset.

Building the model

The first step is to frame the problem: whether it is supervised or unsupervised, and whether it is a classification task, regression task, or something else. The appropriate model is then determined based on the problem type, data size, and computational resources (Géron 2017). All machine learning algorithms were chosen. However, for a more convincing subsequent interpretation of SHAP values, non-additive models are not appropriate.

This study selected tree-based models for the following reasons. First, tree-based models are additive models (Kumar et al. 2020), which are SHAP-friendly. Second, there is no need for standardized data in tree modeling, which can provide a more precise explanation later. Specifically, we selected RF and GBDT. RF has a low variance but high accuracy and may prevent overfitting to an extent by bagging. The RF computation is also small because of the randomization of the samples and features before modeling. The GBDT is appropriate for all classification issues, with good generalization ability and tolerance for different indicators. It is also helpful in handling overfitting, similar to RF. The formulae and steps for these two are as follows:

Random forest (RF)

RF, which combines the results of multiple decision trees using bagging, is a classification and regression technique proposed by Breiman (2001). It performs much better than a single tree (Breiman 1996). In this study, an RF model was developed using an ID3 classification decision tree. The algorithm in our specifications follows the following steps.

(1) Randomly extracted bootstrapped training samples of equal size from the original dataset and arbitrarily selected n(specified constant,$n < N$) feature subsets from all features.
(2) Build a tree model on the dataset developed in step (1). Compute the entropy value for each feature of each sample: $entropy(t) = -\sum_{i=0}^{c-1} p(i|t) log_2 p(i|t)$,[10] and consecutively choose the point to split datasets with the highest entropy value until all the datasets are completely divided (or stop if the pruning condition is met when pruning exists) to form a decision tree.

[9] Negative log-likelihood loss of per sample: $\text{loss}(y, p) = -(y log(p) + (1 - y)\log(1 - p))$. y is true label and $p = \Pr(y = 1)$.

[10] Where t represents the node and i represents the label of the classification, $p(i|t)$ is the ratio of class i instances among the training instances in the ith node. it is frequently used as an impurity measure in machine learning: a set's entropy is zero when it contains instances of only one class (Géron 2017). That is, it is an important indicator to measure the concentration of the sample point distribution. The greater the Entropy is, the more samples can be distinguished, and the more classification information is contained in the features.

(3) Repeat the above steps F times (which can be set manually) to form an RF with F trees.

(4) Vote on the results of F trees and determine the ultimate result using the majority rule. Calculate the classification probability of sample i:$Pr(y_i = 1) = f/F$ and f is the number of decision trees which classify sample i into category one.

Gradient boosting decision tree (GBDT)

GBDT is another tree-based algorithm proposed by Friedman (2001) that works by successively adding predictors to an ensemble, with each one correcting for its predecessor's mistakes. The experimental process for the GBDT in this study is as follows:

We choose log-loss function in this research.

Initialize it to $f_0(x) = \frac{1}{2}log\left(\frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N}(1-y_i)}\right)$, and then repeat steps (1)–(4) for the mth tree ($m = 1, 2, \ldots M$):

(1) Calculate the value of the negative gradient for each sample:$\widetilde{y_i} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)}$ where $L(y_i, f_m(x_i)) = -[y_i log p_i + (1 - y_i) log(1 - p_i)]$ because of the log-loss function and $p_i = \frac{1}{1+e^{[-f_m(x_i)]}}$, simplify it to $L(y_i, f_m(x_i)) = -[y_i f_m(x_i) - log(1 + e^{f_m(x_i)})]$ and the final negative gradient value calculation formula is given as

$$\widetilde{y_i} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)} = y_i - \frac{1}{1+e^{[-(f_{m-1}(x_i))]}} \tag{3}$$

(2) Determine the best regression decision tree using the gradient value computed in step (1), and suppose, in tree modeling, we finally get J terminal nodes: $R_{jm}, j = 1...J$, in which all unions contain all sample sets.

(3) Calculate the output for each terminal node in step (2): $\gamma_{jm} = \frac{\sum_{x_i \in R_{jm}} \widetilde{y_i}}{\sum_{x_i \in R_{jm}} (y_i - \widetilde{y_i}) \times (1 - y_i + \widetilde{y_i})}$.

(4) Update the outcome of GBDT:$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J} \gamma_{jm} I(x \in R_{jm})$, where $I(x \in R_{jm}) = \begin{cases} 1, & x \in R_{jm} \\ 0, & x \notin R_{jm} \end{cases}$.

(5) Obtain the final result by integrating the results of M trees: $F_M(x) = \sum_{m=1}^{M} f_m(x)$, and probability calculation formula of risk occurrence is $p_i = \frac{1}{1+e^{[-F_M(x_i)]}}, i = 1 \ldots N$.

Evaluation of results

In this study, we determined the performance of machine learning models using the receiver operating characteristic (ROC) Curve and the area under curve (AUC) as well as the confusion matrix, which is a multidimensional assessment table of imbalanced samples in classification problems, in which the main indexes are accuracy, recall, precision, and F1-score.[11] Among them, recall can determine the ability of the model to

[11] Specific calculation formulae were shown in the Additional file 2: Appendix S8.

capture the minority class, that is, risk, whereas the F1-score denotes the worldwide quality of the model. Similarly, the ROC and AUC also aid in overall model performance. Furthermore, ROC curves are typically convex. The closer the curve is to the upper left and the larger the AUC area, the better the classification performance for minority categories (Huang et al. 2019). Additionally, the ROC curve can also be applied to raise the recall by shifting the decision threshold, especially during the unbalanced sample classification, and according to experience, the best threshold point is $Max(Recall - FPR)$ (Hilden and Glasziou 1996), which would be used later in model tuning.

Decomposing model results

This step was primarily implemented using the SHAP values. SHAP was suggested by Lundberg and Lee (2017), with visualization in Lundberg et al. (2018). It is suitable for all machine learning model outputs. Nevertheless, it should be known that the optimum case for feature attribution is when the features that are being perturbed are independent at the outset (Kumar et al. 2020; Chen et al. 2020), and for the model, the additive model is the best (Kumar et al. 2020). Therefore, in this study, we attempted to make SHAP more consistent by selecting tree-based models and authenticating the feature subset using the Variance Inflation Factor (VIF).

SHAP is based on the cooperative game theory (Shapley 1953). It explains all features as contributors to the model results. This principle is explained as follows.

For feature $i, i = 1, 2, 3 \ldots N$, if the contribution $\psi_i(N, v)$, i.e., Shapley value satisfies the properties of Efficiency, Symmetry, Dummy, and Additivity,[12] then its precise calculation formula is expressed as follows:

$$\psi_i(N, v) = \frac{1}{N!} \sum_{S \subseteq N \setminus \{i\}} |S|!(|N| - |S| - 1)![v(S \cup \{i\}) - v(S)] \tag{4}$$

where $v(\cdot)$ represents the sum of the contributions made by all features in brackets.

Denote the Shapley value of feature i and sample $j, j = 1, 2, 3 \ldots M$ as ${\psi_i}^j(N, v)$, work out the Base Value $E(\hat{y})$, which is the average of the fitted values of the target variable in the training set, then the projected value of sample j can be expressed as follows:

$$y_j = E(\hat{y}) + {\psi_1}^j(N, v) + {\psi_2}^j(N, v) + \cdots + {\psi_N}^j(N, v) \tag{5}$$

where the magnitude of ${\psi_i}^j(N, v)$ represents how much contribution the feature i does to sample j, and the sign shows its impact direction.

Lundberg and Lee (2017) proposed several improved calculation methods suitable for complex models. Although the original study summarized the characteristics of local accuracy, missingness, and consistency and indicated the different calculation procedures, we do not go into detail in this study because the basic principle is the same as above. Overall, an appropriate SHAP kernel can be chosen based on the selected model: Kernel SHAP, which is model-agnostic (Lundberg and Lee 2017), Tree SHAP for high-speed and precise computation of tree-based models (Lundberg et al. 2020), Deep SHAP and DeepLIFT for deep learning models (Shrikumar et al. 2017).

[12] A description of the specific characteristics can be found in Additional file 2: Appendix S8.

Statistical inference on the decomposition

This step was performed using the Shapley regression to evaluate the statistical significance of the predictors according to the SHAP calculated in the previous step. It was first suggested by Joseph (2019), involving calculated the Shapley Share Coefficients (SSC) and regression.

Regression

The Shapley regression is the regression of the model output on the SHAP values of individual data points in a limited local region. Joseph (2019) experimented using numerical cases and Brazilian educational datasets. Suss and Treitel (2019) and Bluwstein et al. (2020) have proposed extensions. The construction and explanation are summarized as follows.

Let the dependent variable in the regression be $ln\left(\frac{y_{risk}}{1-y_{risk}}\right)$, then the Shapley regression model is expressed as follows:

$$ln\left(\frac{y_{risk}}{1-y_{risk}}\right) = \beta_0 + \beta_1\varphi_1 + \beta_2\varphi_2 + \cdots + \beta_n\varphi_n \tag{6}$$

Theoretically, the coefficients determine the alignment of Shapley components φ_i with the target y_{risk}. Coefficient values near one indicate the optimum alignment and convergence of the learning process. A value greater than 1 implies that the model underestimates the impact of the variable on the outcome, with a value less than 1 signifying the opposite (Buckmann and Joseph 2022). A coefficient of zero indicates no clear alignment. Negative coefficients indicate a poor fit for the training model. However, in the case of out-of-sample predictions, they point to surprising findings, especially those with high significance (Joseph et al. 2021). In terms of specific values, suppose φ_1 increases by one unit and other variables remain unchanged. The new regression equation can be transformed into

$$\begin{aligned}\ln\left(\frac{y_{risk}}{1-y_{risk}}\right)' &= \beta_0 + \beta_1(\varphi_1 + 1) + \beta_2\varphi_2 + \cdots + \beta_n\varphi_n \\ &= \beta_0 + \beta_1\varphi_1 + \beta_2\varphi_2 + \cdots + \beta_n\varphi_n + \beta_1 = \ln\left(\frac{y_{risk}}{1-y_{risk}}\right) + \beta_1\end{aligned} \tag{7}$$

Take the natural logarithm on both sides of the equation as:

$$\left(\frac{y_{risk}}{1-y_{risk}}\right)' = \left(\frac{y_{risk}}{1-y_{risk}}\right) \cdot e^{\beta_1} \tag{8}$$

It can be observed that $\left(\frac{y_{risk}}{1-y_{risk}}\right)$ will become e^{β_1} times the original. That is, the probability ratio will vary by $100 \times \left(e^{\beta_1} - 1\right)\%$, consistent with Suss and Treitel (2019).

Shapley share coefficients (SSC)

Define the SSC of variable x_k:

$$\Gamma_k^s\left(\hat{f}, \Omega\right) \equiv \left[sign\left(\hat{\beta}_k^{lin}\right)\left\langle\frac{\left|\varphi_k^s\left(\hat{f}\right)\right|}{\sum_{l=1}^{n}\left|\varphi_l^s\left(\hat{f}\right)\right|}\right\rangle_{\Omega_k}\right]^{(*)} \in (-1, 1) \tag{9}$$

where $\varphi_k^s\left(\widehat{f}\right)$ denotes the SHAP value of the variable x_k, $\langle\cdot\rangle_\Omega$ represents the mean of all data in Ω, $sign\left(\widehat{\beta}_k^{lin}\right)$ stands for the sign of the variable x_k in the linear model,[13] and $(*)$ shows the significance level against the null hypothesis.

The $\Gamma_k^s\left(\widehat{f},\Omega\right)$ denotes the expected change in probability caused by variable x_k, relative to other variables. Moreover, as $\Gamma_k^s\left(\widehat{f},\Omega\right)$ has been normalized, the sum of the absolute values of SSC is 1.

It should be known that Shapley regression also has certain practical limitations, such as the necessity of dividing the data into a training set and test set (standard in machine learning) and the necessity of using CV (to help model's convergence), the details can be found in Joseph (2019).

Workflow performance

Feature engineering

The database contains 47 variables. Therefore, first, we eradicated worthless variables (such as "xx_ipolated" which denotes whether to interpolate the xx feature) and those with numerous missing values, finally leaving 22 original variables. Table 2 lists the indicators and their categories.

Feature construction

The features constructed according to economic meaning are listed in Table 3. They were classified into the seven categories mentioned above.

After eliminating the time trends of the features mentioned above, 63 features were obtained. Taking the US as an example, plotting each indicator's value over time shows that the time trend has been roughly removed (see the graph presented in Additional file 2: Appendix S3). We maintained all indicators for future selection because it is uncertain how the indicators would be more beneficial to the forecast. So far, 63 indicators have been prepared for this purpose.

Feature selection and evaluation

MI, Spearman correlation matrix, and RFE were used for feature filtering and selection.

(1) MI filtering: Compute the mutual information scores of each feature and take the average of 100 iterations to prevent computing contingency, and then remove features whose MI scores are less than 0 (the remaining 50 indicators and their MI scores are presented in Additional file 2: Appendix S2).
(2) Spearman correlation matrix filtering: Compute the Spearman correlation matrix (shown in Additional file 1: Accessory 1) for the residual features following step (1), determine the set of features with a coefficient greater than 0.7,[14] and eliminate the feature with a smaller MI score.

[13] For data of China, we extracted it from the annual mean of this feature's SHAP value.

[14] As a rule of thumb, a correlation coefficient between two characteristics bigger than 0.7 indicates a strong correlation between the two indicators.

Table 4 Final features

Variables	Symbols	Construction method
Global net export	GlobalNetExport	Original value
Domestic net export	NetExport	(Exports — imports)/GDP
Narrow money	NarrowMoney	Narrow money/GDP
Domestic yield curve slope	YieldCurveSlope	Short-term interest rates–Long-term interest rates
National financial concentration	FinancialConcentration	Government revenue/GDP
Loan cost	CreditCostGrowth	Growth rate of loans*Long-term interest rates
Domestic credit	Credit	Total credit/GDP
Domestic credit annual increase	Credit-Dif	(Total credit/GDP)[t] — (Total credit/GDP)[t–1]
CPI annual increase	CPI-Dif	CPI[t] — CPI[t–1]
Dividend yield ratio	DidYield	Dividend / Share price

(3) RFE: This method combines RFE and RF. Two features were removed each time, and the final feature subset was selected by the highest score using the mean negative log-likelihood loss of fivefold CV with the remaining features.

There are 38 features left after steps (1) and (2). We called the RFE and plotted the number of features remaining following each elimination and their respective performances (see Additional file 2: Appendix S2), with 10 features performing the best, signifying that this feature subset is the most efficient for risk prediction. The characteristics and their symbols are listed in Table 4.

The following is the link between each indicator and systemic financial crisis, as well as the assumptions of its impact on risk:

The systemic financial crisis can spread through international trade channels. On the one hand, a country's international trade influences the operation of enterprises, which affects enterprises' financing and credit and has an impact on commercial banks and the financial system. On the other hand, international trade also affects the banking and financial systems through cross-border transactions and international settlement business (Yang and Wang 2021; Asgharian et al. 2013). NetExport represents the degree of dependence on foreign trade, and the more closely related it is to the international economy, the more likely it is to be affected by spillovers from foreign systemic financial risk. GlobalNetExport indicates the global trade environment, which reflects the vulnerability of global countries' economic and financial systems. The indicator represents the imbalance of the international trade market and could affect the national financial system; a larger GlobalNetExport indicates that trade deficits are concentrated in a small number of countries, and the larger the value, the worse the deficits, and therefore more prone to sovereign debt risks and spread to other countries, which could result in financial turmoil; a small GlobalNetExport shows a slowdown or even recession in the international economy, which increases the systemic financial crisis spillover. Therefore, we assume that when the feature value is moderate, the possibility of a crisis is low.

NarrowMoney refers to the amount of money in circulation and the current purchasing power. It also represents market demand and can affect systemic financial crises through bank credit channels. Demand expansion stimulates investment by enterprises. When demand expands, consumption and investment are strong and enterprises and

households have a high willingness to borrow and a strong ability to repay loans, which have little negative impact on the systemic financial crisis. Demand retrenchment leads to poor-quality credit assets, threatening the stability of the entire banking industry and financial system (Wang and Tian 2016). At the same time, liquidity may be transferred to the financial system if investment in the real economy is nonprofitable owing to insufficient demand, which may also cause greater pressure on the financial system and make it more prone to risks. In summary, demand expansion is more likely to lead to future prosperity and less market volatility than insufficient demand. Therefore, our assumption is that the lower the feature value, the more likely the risk.

The YieldCurveSlope represents market expectations or confidence. When market expectations change, investors' investment behaviors change accordingly, which leads to market volatility and risk spillovers and is not conducive to the stability of the financial system (Yang and Wang 2021). Specifically, pessimistic market expectations prompt investors to sell their financial assets, resulting in a decline in asset prices, which further leads to worse market expectations and larger financial asset selling and may even trigger the "herd effect." In other words, asset prices and market expectations form a vicious spiral cycle with self-reinforcing characteristics, driving financial system risk. However, overly optimistic market expectations could lead to an excessive expansion of credit and leverage, exacerbating the vulnerability and accumulating risk of the financial system (Ma 2013). In contrast, the vicious circle created by pessimistic expectations is more likely to break the financial system. In this study, YieldCurveSlope is calculated by short-term interest rates minus long-term interest rates, such that larger value indicates stronger pessimistic expectations. Therefore, our assumption is that the higher the slope, the more likely the risk.

FinancialConcentration is a fiscal indicator. Low government revenue stimulates government borrowing. The expansion of government credit could crowd out the total credit scale and cause a resource mismatch, which is not conducive to financial system stability. Moreover, if the government fails to repay its debts and the government's default rate rises, it will influence the quality of financial assets and lead to a systemic financial crisis. In other words, there is the possibility of a fiscal crisis transforming into systemic risk (Xiong and Jin 2018). For large government revenue, the government is less likely to face debt risk. However, if revenue is increased by raising taxes, it could reduce corporate profits, and companies are more likely to expand production through loans as well as pressure the banking and financial systems (Mao et al. 2018). Considering that taxes are a part of government revenue, we speculate that small government revenue has a greater impact on risk. Therefore, we assume that the higher the revenue, the less likely the risk.

Among the credit indicators, CreditCostGrowth represents the cost of long-term loans, Credit is the scale of credit, and Credit-Dif is the annual increase in the credit scale. These indicators measure credit security from different perspectives. Financial turbulence typically occurs during loan expansions. Massive lending or rapid credit increases may reduce bank credit ratings and lead to bankruptcy, which is directly related to the risk of banking and financial systems (Schularick and Taylor 2012). Moreover, the cost of loans also measures the fragility of credit repayment: the longer the credit repayment period, the more unstable the factors affecting the repayment,

Table 5 T-test

Variables	Mean of non-risk period	Standard deviation for non-risk periods	Mean of risk period	Standard deviation for risk periods	T-test and significance	*P* value	VIF
DidYield	0.04	0.02	0.04	0.02	−2.00**	0.048	1.46
Credit	64.12	31.78	80.92	37.05	−4.20***	0.000	1.57
CPI-Dif	178.12	188.60	190.20	234.54	−0.48	0.633	1.58
Credit-Dif	1.06	3.35	3.42	5.31	−4.17***	0.000	1.29
CreditCost-Growth	8.51	19.01	17.52	17.65	−4.38***	0.000	1.36
NarrowMoney	21.38	17.12	19.98	15.25	0.76	0.447	1.31
NetExport	−0.96	5.06	−3.00	7.07	2.69***	0.008	1.25
FinancialCon-centration	19.87	10.18	16.72	11.11	2.82***	0.005	1.76
YieldCurveSlope	−0.83	1.72	0.21	1.42	−5.58***	0.000	1.15
GlobalNetExport	−1.48	2.49	−1.84	3.21	1.02	0.309	3.28

The null hypothesis of the t-test is that the mean value of the difference between the risk and non-risk periods is 0; ***, **, and * indicate that the null hypothesis is rejected at significance levels of 1%, 5%, and 10%, respectively

such as personal income and interest rate, adding pressure on banks' balance sheets (Bluwstein et al. 2020). Consequently, we assume that for all three factors, the higher the value, the more likely the risk.

CPI-Dif is the yearly increase in inflation reflecting the degree of currency devaluation. On the one hand, the rise in inflation could reduce enterprises' profits, which may lead to enterprises' repayment default or additional loans and affect the systemic financial crisis through credit channels. Similarly, deflation also affects the financial system through this channel, as deflation adds a burden to enterprises' debt, and when enterprises are unable to repay, rising nonperforming loans would increase risk (Wei and Yang 2017). On the other hand, deflation, or a continuous rise in inflation, raises risk by lowering market expectations (Ma 2013). Taken together, we propose the assumption that the more the value deviates from 0, the more likely the risk.

DidYield refers to dividend distribution and investment returns. It affects the enterprise's profit and thus influences its investment and debt. Enterprise investment has a direct impact on the macroeconomy, affecting the stability of the financial system, whereas the scale and quality of enterprise loans affect bank credit and systemic financial crises. In addition, DidYield is associated with stock market returns and can be used to measure market prosperity. A booming stock market can create bubbles and risk spillover. In conclusion, we assume that the higher the dividend rate, the more likely the risk (Zhou et al. 2020).

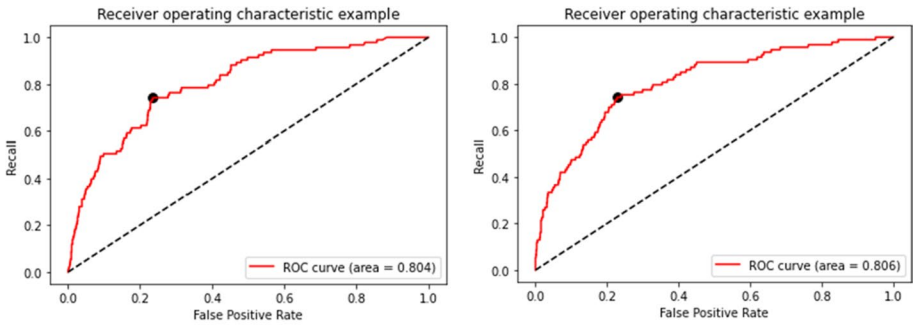
Thus far, a dataset of 1005 samples from 16 nations and 10 features has been obtained. It is fed into the model, uses CV and Bayesian Optimization for training, and performs out-of-sample measurements with Chinese systemic risk. For clarity, owing to data limitations, only the risk during 1990–2020 was analyzed.

Data description

T-test results for the final features are presented in Table 5, which shows that most features are significantly different between the risk and non-risk periods, except for

Table 6 Model performance in training set

Algorithms	AUC (CVmean)	Accuracy	Weighted precision	Risk recall	Weighted F1-score
RF	0.804	0.76	0.90	0.73	0.81
GBDT	0.806	0.77	0.90	0.73	0.81



a Based on RF **b** Based on GBDT

Fig. 2 ROC graph in training. The dots in the figure indicate the threshold points used for tuning

CPI-Dif, NarrowMoney, and GlobalNetExport. We also computed the VIF among the selected features, which were all less than 5. Based on this rule of thumb, there was no multicollinearity (Jin and Tao 2016).

Building the model

To build the model, we needed to determine the hyperparameters of each model, which were determined by Bayesian Optimization and fivefold cross-validation.

For the RF, we experimented with six hyperparameters. Among them, "criterion" and "n_estimators" are the roots of RF, and we finally settled on 460 trees with Entropy to form a random forest. The parameters "max_depth," "min_samples_leaf," "min_samples_split" and "max_features" were designed to prevent over-fitting of the model, and we finally determined the parameters as 12, 1, 2, 2. The parament "class_weight" adjusts the sample imbalance, and we determined the ratio of non-risk samples to risk samples as 1:1.225, and the final AUC (CVmean) can reach 0.804.

For the GBDT, we experimented with seven hyperparameters. We first plotted the learning curve of "learning_rate" and "n_estimators" to determine the approximate range. We then employed the grid search method to pinpoint their values, where "learning_rate" was 0.061 and "n_estimators" was 40. On this basis, we drew the learning curve of the remaining six parameters to determine the parameter range and fed them into the Bayesian Optimization. The final parameter values are: "subsample" is 0.7744261168207747, "min_samples_split" is 8, "min_samples_leaf" is 2, "max_depth" is 4, and "max_features" is 3. The final AUC (CVmean) was 0.806.

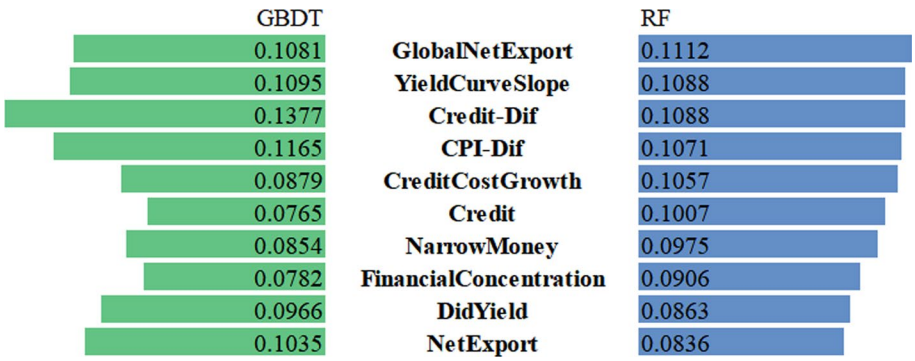


Fig. 3 Variable importance

Prediction of the model's performance

The confusion matrix, ROC, and AUC measure the prediction of the model's performance. The findings for the training set are summarized in Table 6. We also present the ROC curves of 5-folds cross-validations in the training set in Fig. 2.

None of the trained models suffered from severe overfitting. The Weighted precision, risk recall and Weighted f1-score are exactly the same. The accuracy of the GBDT was higher in the training set, and the AUC of the GBDT was also slightly higher, as shown in Fig. 2. Overall, there was little difference in the performances of the two models. Although the GBDT's accuracy is slightly higher, the RF models lower the variance when modeling, which makes the RF model more stable. Therefore, we chose RF for the interpretation of the latter. The GBDT findings are shown in the Additional file 2: Appendix.

Variable importance

Variable importance analysis was applied to identify the essential variables. However, this is only applicable to the overall training sample and cannot be conducted for the out-of-sample set, which restricts our research to China. Therefore, we only present it in Fig. 3 as a supplement and compare it with the subsequent analysis. This finding highlights the significance of GlobalNetExport, Yield CurveSlope, Credit-Dif, CPI-Dif, and CreditCostGrowth in terms of global systemic risk.

SHAP value

We computed the contribution of each feature to the systemic financial crisis for each year and analyzed it independently from the time and feature dimensions.

Time dimension

For each indicator, we averaged the SHAP value of all countries in each year, and the average SHAP value of each indicator in each year indicated the contribution of the indicator to the global average risk probability in that year.[15]

This section describes the changes in risk drivers in the time dimension of the training set. However, because the training set had a wide time range and was not the focus

[15] We also plotted the average SHAP of both RF and GBDT in Additional file 2: Appendix S11.

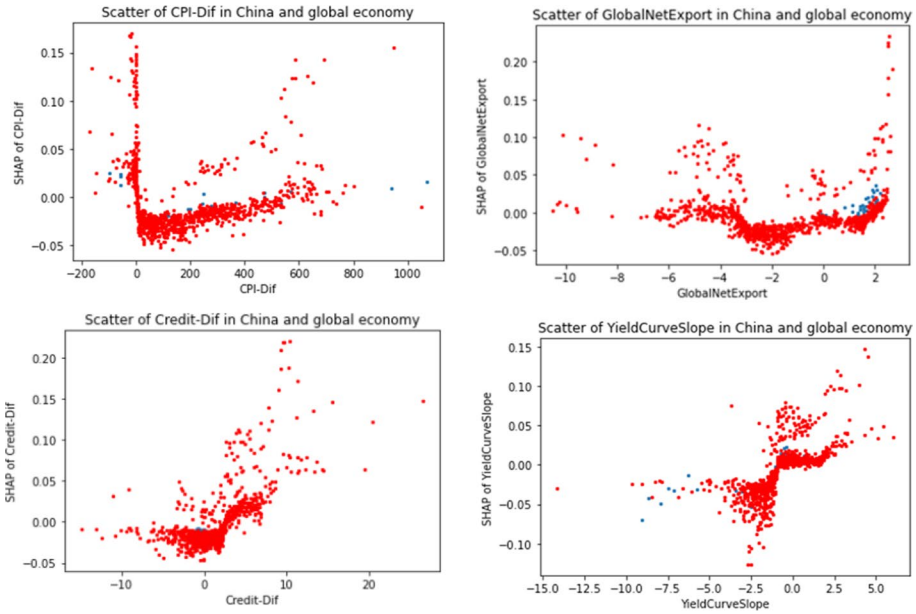


Fig. 4 Scatter plots showing the relationship between SHAP value and feature value. Only a few of them are shown here. The others can be found in Additional file 2: Appendix S4. Red dots represent the global training data points, while blue dots are data from China

Table 7 Threshold effect

Features	Threshold	Notes
CreditCostGrowth	7	Larger feature value, larger SHAP value, larger impact; Feature value beyond the threshold, the pulling force increases sharply
Credit	30(100)	
Credit-Dif	2.5	
YieldCurveSlope	− 1	

The brackets 30(100) indicate that the probability of risk rises sharply when the feature value exceeds 30, and if it exceeds 100, it is almost impossible not to occur the risk

of this study, we only analyzed the three periods with the top three largest SHAP values. In 1877, the average SHAP of FinancialConcentration and CPI-Dif were 0.097 and 0.094, respectively, maximally increasing the risk. In 1929–1931, DidYield and CPI-Dif made the greatest contributions, boosting the risk probability by 0.095 and 0.103 in 1930, respectively. From 2006 to 2008, the contributions of most features to risk increased, particularly CreditCostGrowth, GlobalNetExport, and Credit. GlobalNetExport made the largest contribution in 2006, with an SHAP of 0.090, and in 2007, CreditCostGrowth and Credit made the most significant contributions to the risk probability, with SHAP of 0.100 and 0.074, respectively.

Feature dimension

In the feature dimension, as Fig. 4 demonstrated, we drew scatter plots to reveal the relationship between the SHAP values and the corresponding feature values.[16] Most features have a non-linear effect on risk, which can be explicitly divided into threshold

[16] GBDT's plots showed in Additional file 2: Appendix S6.

Table 8 Interval effect

Features	Safe-haven zone	Notes
FinancialConcentration	[13, 35]	Concave impact; Compared with the feature value larger than the maximum of the safe-haven zone, the feature value less than the minimum of the safe-haven zone is more likely to increase the probability of risk
CPI-Dif	[3, 200(400)]	
NarrowMoney	[7, 30]	
NetExport	[2(0), 7]	
DidYield	[0. 01, 0. 07]	Concave impact; Compared with the feature value less than the minimum of the safe-haven zone, the feature value larger than the maximum of the safe-haven zone is more likely to increase the probability of risk
GlobalNetExport	[− 3.5, 2]	

The brackets 200(400) mean the same thing as above. The brackets 2(0) indicate that the safest range is [2,7], and the worst ratio cannot be less than 0

and interval effects. Finally, we summarized the thresholds and intervals of the features based on the global training set, as presented in Table 7 and 8, respectively.

Risk is more likely to occur when the yield curve is inverted. In fact, when the difference between the short-term interest rate and the long-term interest rate is close to -1, it is already a sign of risk. If the difference continues to increase, the risk continues to increase with increasing pulling force. As shown in the lower right of Fig. 4, the SHAP value of YieldCurveSlope, which is the impact of the slope on crisis, shows an overall upward trend in the scatter plot, consistent with the previous assumption. Specifically, as the slope increases, its effect on reducing the risk gradually decreases until the value of − 1, it turns to increasing risk, which indicates a significant threshold effect of the yield curve slope on the risk.

The credit indicators are similar and their SHAP values show an upward trend. Moreover, financial risk is more likely to occur when annual inrease of credit jumps by more than 2.5% in a year. The higher the annual increase, the greater the risk probability. This indicator can increase risk probability by 20% within a year. The Credit to GDP ratio and growth in credit costs also show thresholds of 30% and 7%, respectively. All these findings are consistent with previous assumptions.

NetExport has a V-shaped effect on risk; when the ratio falls to the left of 4, the driving force of risk decreases as the ratio increases. When the ratio exceeded 4, the force started to increase. Moreover, when the ratio varied from 0 to 2%, it had a modest impact on the risk, increasing in only a few cases. However, the safest ratio ranged from 2 to 7%, which barely increase the risk probability. In other words, it is relatively safe for a country to maintain a proper trade surplus or at least no deficit, and a deficit is more likely to cause risks than a surplus. GlobalNetExport's scatter plot tends to be U-shaped and is higher on the right than on the left (top right of Fig. 4). Specifically, the average ratio of net exports to GDP of all countries in the world exceeds 2% or is less than − 3.5%, both of which increase risks, with the former being more powerful. Combined with the previous analysis and assumptions, we conclude that the sovereign debt risk contagion caused by the deficit of a few countries has a greater impact on the financial system than the global economic slowdown.

Table 9 RF-Shapley regression in global set (Sorted by the size of share)

Feature	Sign	Share (SSCvalue)	Coefficients	*P* value	Exponentiation	Change of $\frac{y_{risk}}{1-y_{risk}}$
CreditCostGrowth	+	0.1379***	0.0401	0.000	1.0409	Significantly increased by 4.09%
CPI-Dif	+	0.1365***	0.0373	0.000	1.0380	Significantly increased by 3.80%
YieldCurveSlope	+	0.1222***	0.0469	0.000	1.0480	Significantly increased by 4.80%
Credit-Dif	+	0.1151***	0.0547	0.000	1.0562	Significantly increased by 5.62%
GlobalNetExport	−	0.1013***	0.0323	0.000	1.0328	Significantly decreased by 3.28%
NarrowMoney	−	0.0970	-0.0455	0.000	0.9555	Insignificant increase
FinancialConcentration	−	0.0959	-0.0157	0.038	0.9844	Insignificant increase
Credit	+	0.0816***	0.0560	0.000	1.0576	Significantly increased by 5.76%
DidYield	+	0.0610	-0.0255	0.000	0.9748	Insignificant increase
NetExport	−	0.0515	-0.0393	0.000	0.9615	Insignificant increase

***, **, and * respectively means that the null hypothesis is rejected at the significance level of 1%, 5%, 10%

The figure of CPI-Dif is in line with previous analyses and shows that the driving force of inflation on risk is concave, first decreasing and then increasing (as shown in the top left of Fig. 4). Moreover, deflation has a more pronounced impact on risk than does inflation. The impact of NarrowMoney on risk has a shape similar to that of the CPI-Dif, which is also consistent with the previous assumption. Moreover, risk is more likely to increase when NarrowMoney is low, especially when it is lower than 7.

For fiscal indicators, if the FinancialConcentration is too small (e.g., lower than 13%), it is easy to accumulate the risk of government debt default. However, if it is too large (e.g., more than 35%), it is not conducive to the healthy operation of the entire economy, resulting in the accumulation of risk. Both can damage the stability of the financial system, but the force of the latter is much lower than that of the former. Therefore, overall, lower government revenue is more prone to risk, which is consistent with previous analyses and assumptions. Combining the previous analysis, we can conclude that FinancialConcentration is more prone to risk through government debt defaults.

For DidYield, as expected, an excessive ratio (e.g., greater than 0.07) leads to overheated investment and bubbles and increases risk. However, beyond our assumption, a small ratio also increases the risk. This may be because the very low rate is closely related to the incomplete development of the stock market. In such cases, it is difficult for enterprises to finance, which results in a single debt structure and puts pressure on the banking industry. However, its impact was far less than that of an excessive rate. According to the plot, the optimal range of the market dividend rate is 0.01–0.07.

We also drew PDP plots to verify robustness, and the plots are shown in Additional file 2: Appendix S7. These trends were primarily consistent with those of SHAP.

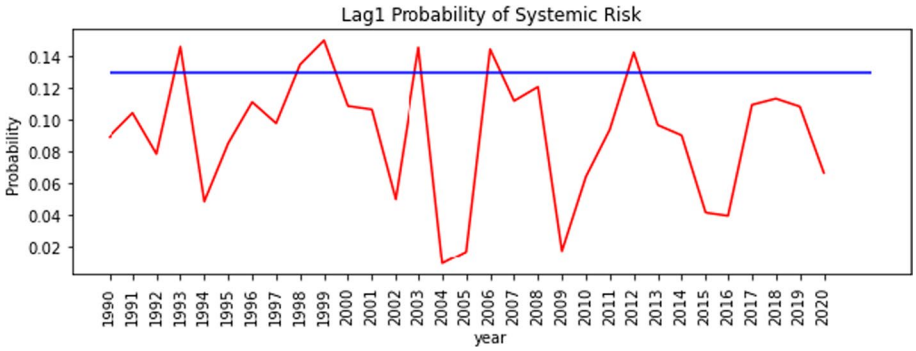


Fig. 5 Model-calculated probability of systemic financial risk in China

Shapley regression

In this section, we turn to statistical inference using the Shapley regression to assess the confidence of the prediction. The results[17] of the training set are summarized in Table 9.

Table 9 shows that CreditCostGrowth, YieldCurveSlope, CPI-Dif, Credit-Dif, GlobalNetExport, and Credit are statistically significant in global dataset. Other indicators' changes are not markedly aligned with changes in the risk probability (Joseph et al. 2021). We speculate that the possible reason for this is that we did not fit the global dataset completely in order to improve the out-of-sample predictive and generalization capabilities of the model.

In particular, the coefficient of Credit shows that an increase of one unit in its SHAP value will increase $\frac{y_{risk}}{1-y_{risk}}$ by 5.76%, showing its significant role in predicting systemic financial risk in the global dataset. Another factor that significantly increases the risk probability is Credit-Dif, where an increase in the SHAP value by one unit leads to an $\frac{y_{risk}}{1-y_{risk}}$ increase of 5.62%.

Similarly, for CreditCostGrowth, a one-unit increase in SHAP raises $\frac{y_{risk}}{1-y_{risk}}$ by 4.09%. YieldCurveSlope and CPI-Dif were also vital predictors of risk, with a unit change in the SHAP value boosting the probability ratio by approximately 4%.

Additionally, a one-unit increase in the SHAP of GlobalNetExport decreased the probability ratio by 3.28%. As previously noted, GlobalNetExport had a U-shaped impact. Although the impact of GlobalNetExport to risk is going to jump when it exceeds 2, it is an uncommon circumstance. Extreme values are expected to be ignored when the OLS is fitted. Therefore, the negative coefficient of GlobalNetExport indicates that it contributes less to the systemic financial crisis as the economy accelerates out of a recession.

Trained workflow in China's risk

Building the model

We used the established model to measure systemic financial risk in China. Similarly, RF and GBDT captured six periods of risk rising since 1990 in China. We drew a time plot of the model-calculated probability of systemic financial risk in China after 1990, as presented in Fig. 5, where the blue horizontal line denotes the optimal threshold in tuning

[17] GBDT's table showed in Additional file 2: Appendix S9.

with the ROC, and the probability falls in the area above the blue line which shows the risk. The graph is based on the RF results, and the GBDT line is similar; therefore, it will not be repeated (see Additional file 2: Appendix S5).

The following is an analysis of the timing change[18] of systemic risk probability in China from a macroeconomic perspective, based on China's reality and the relevant literature.

In 1993, the probability of a systemic financial crisis in China increased for the first time since 1990. Because China's credit growth increased in this year due to Deng Xiaoping's Southern Talks in 1992, which affected the stability of the financial system (Chen et al. 2009; Zhang 2015).

From 1996 to 2000, the Southeast Asian financial crisis significantly affected China's foreign trade and stock markets and led to deflation (Wei and Yang 2017), which aggravated the turmoil in China's financial system. In addition, China's tax reform in 1994 created fluctuations in government revenue during this period (Chen et al. 2009), which also increased the risk probability. However, benefiting from China's proactive fiscal and monetary policies, the probability of a systemic financial crisis ultimately decreased (Wei and Yang 2017).

By 2003, there was another increase in China's risk because during this time, some large state-owned banks had huge burdens of nonperforming loans and some were even on the verge of "technical bankruptcy" (e.g., the Agricultural Bank of China; Li and Liang 2021; Zhang 2015). The real economy was hit by the outbreak of SARS, which lowered market expectations and affected the financial system (Yang et al. 2020).

In 2006, China increased the release of loans, leading to an overheating of investments in fixed assets (Chen et al. 2009). The U.S. subprime crisis occurred in 2007. These factors jointly triggered fluctuations in China's financial system and increased the probability of systemic financial crises in 2006 and 2007[19] (Li and Liang 2021; Tao and Zhu 2016; Guo et al. 2018). By 2008, owing to the Four Thousand Billion Stimulus Plan, the risk probability had not increased significantly (Guo et al. 2018; Li and Liang 2021).

From 2011 to 2013, there was a liquidity strain in China's financial system and interbank market, along with the European sovereign debt crisis and recession expectations in global economies, resulting in another peak in risk probability (Guo et al. 2018; Li and Liang 2021). However, as revealed by Li and Liang (2021), the impact of the liquidity squeeze on risk was not durable, and risk probability decreased in 2014 and 2015.

From 2016 to 2017, China experienced a surge in credit again: on the one hand, China's "steady growth" policy led to an increase in the supply of medium-term and long-term loans; on the other hand, real estate loans also grew significantly because of the rapid development of real estate and the rise of housing prices. Thus, the probability of the systemic financial crisis increased again (Guo et al. 2018; Li and Liang 2021). In addition, China's efforts to deleverage its financial sector during this period were also an important factor in raising risks. Moreover, the outbreak of a trade conflict between China and the United

[18] It should be noted that China officially proposed the goal of establishing a socialist market economy system in 1992, and the market was not well established before 1992, which may lead to imprecise outcomes, therefore, we started our analysis in 1992. Meanwhile, the outbreak of COVID-19 pandemic in 2020 may influence the systemic risk, but we do not include epidemiological factors in our model, so the risk in 2020 is also excluded from the analysis. And we shall consider the non-economic factors in future studies.

[19] This conclusion has not reached a consensus in the academia and some papers suggest that there are no risks in China during this time, such as Zhang (2015).

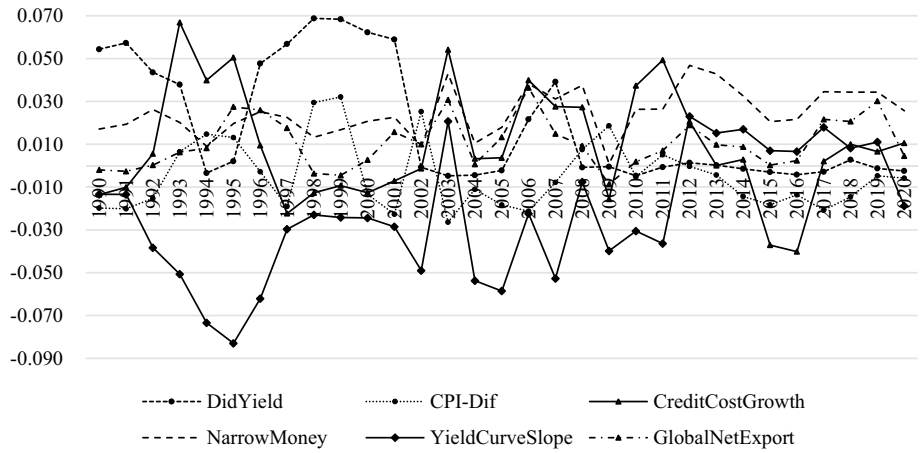


Fig. 6 Annual SHAP value trend of some variables with China data

States in 2018 also impacted market expectations and international trade, driving up risk probability.

In summary, China's financial system was more prone to risk during 1993, 1996–2000, 2003, 2006–2009, 2011–2013, and 2017–2019, which is in line with the actual conditions and most of the literature. In other words, the trained model can simulate the general trend of risk in China, which proves the validity of the model on Chinese data and demonstrates its good generalization.

SHAP value

We analyzed the results of SHAP decomposition in two dimensions, as we did previously.

Time dimension

Regarding the time dimension, Fig. 6 displays the annual SHAP value trends of the variables with apparent fluctuations in China.[20] For example, according to Fig. 6, the significant drivers of each increase in systemic risk in China over the past 30 years[21] are as follows:

Form 1992 to 1993, the driving force of CreditCostGrowth increased sharply, consistent with the previous timing analysis. By 1994, although CreditCostGrowth dropped, it was still at a high level, and its driving force for risk (0.040) ranked first. In addition, the CPI-Dif also made a significant contribution (0.015), and its contribution rose the most in this year, in line with the reality of inflation. Moreover, DidYield was the most prominent driving force preceding 1993, which may be because China's DidYield was extremely low during this period, and according to our previous link analysis it could burden the banking and financial sectors. Overall, before 1995, DidYield and CreditCostGrowth were the most prominent driving forces, and CreditCostGrowth and CPI-Dif were the most sensitive to risk. Furthermore, compared to CPI-Dif, CreditCostGrowth and DidYield exhibited a larger range of changes.

[20] GBDT'S plot showed in Additional file 2: Appendix S12.

[21] Referring to footnote 18, scenarios before 1992 and after 2020 were still not included in the analysis here.

From 1996 to 2000, DidYield was the most influential factor in increasing the risk probability, and its effect peaked in 1998 and 1999. This can be attributed to the impact of the Southeast Asian financial crisis on trade, the real economy, and the accompanying deflation, which led to poor corporate earnings, reduced dividends, and fluctuating stock markets. GlobalNetExport, NarrowMoney and CPI also have direct driving forces on risk. In addition, FinancialConcentration had a notable impact on risk between 1994 and 1999, which is in line with the analysis of tax reform mentioned above.

In 2003, CreditCostGrowth contributes the most to risk. The volatility of CreditCostGrowth may have been due to China's chaotic banking system during this period, with a low capital adequacy ratio, poor risk management, and poor operating performance. In 2002, SARS broke out, and both disorganized banking system and SARS could lower market expectations (i.e., YieldCurveSlope). SARS may have also negatively affected risk probability through the international trade channel (i.e., GlobalNetExport). In addition, NarrowMoney's value increased in China during this time, but the risk also increased, which is not consistent with the theoretical analysis and conclusions of the training set that an increase in NarrowMoney is generally related to the expansion of consumption and demand with a small impact on the financial system. This may be because in China, on the one hand, enterprise demand deposits account for a high proportion in NarrowMoney and the rise in NarrowMoney is mainly driven by the rapid increase of enterprise demand deposits, which reflects in the falling of return on assets, the decrease of enterprise investment intention, and the decline in investment in reality. On the other hand, Zhou and Wang (2009) pointed out that money supply is closely linked to fluctuations in real estate asset prices, so the rise in real estate asset prices could lead to an increase in real estate demand deposits, which then caused the growth of NarrowMoney. Therefore, the expansion of NarrowMoney in China may result in a decrease in investment and an increase in real estate deposits, which is distinctive and not conducive to the development of the real economy and financial system (He and Yu 2018). Thus, the surge in China's property prices in 2003 can explain why NarrowMoney's expansion led to a rise in risk.

During 2006–2009, CreditCostGrowth was the largest risk driver in 2006, in line with China's 2006 credit boom. By 2007, it had changed to DidYield, possibly because the subprime crisis led companies to take protective measures to cut dividends, which rebounded in 2008 after the Chinese government took effective measures. GlobalNetExport represents an unfriendly international environment that also makes a few contributions. In addition, the contributions of NetExport and CPI-Dif increased to different degrees, which corresponded to the impact of external demand and deflation during this period. Furthermore, GlobalNetExport has a more significant impact on risk than NetExport.

From 2011 to 2013, the "cash crunch" increased banking risk and lowered market confidence. CreditCostGrowth, GlobalNetExport, YieldCurveSlope, and NarrowMoney contributed significantly to risk, which is related to the poor global environment and poor expectations caused by the European debt crisis in 2012 as well as the rise in China's real estate asset prices in 2013.

From 2016 to 2017, CreditCostGrowth increased dramatically, making it the most sensitive risk driver; that is, its SHAP increased by 0.042 from 2016 to 2017, which is

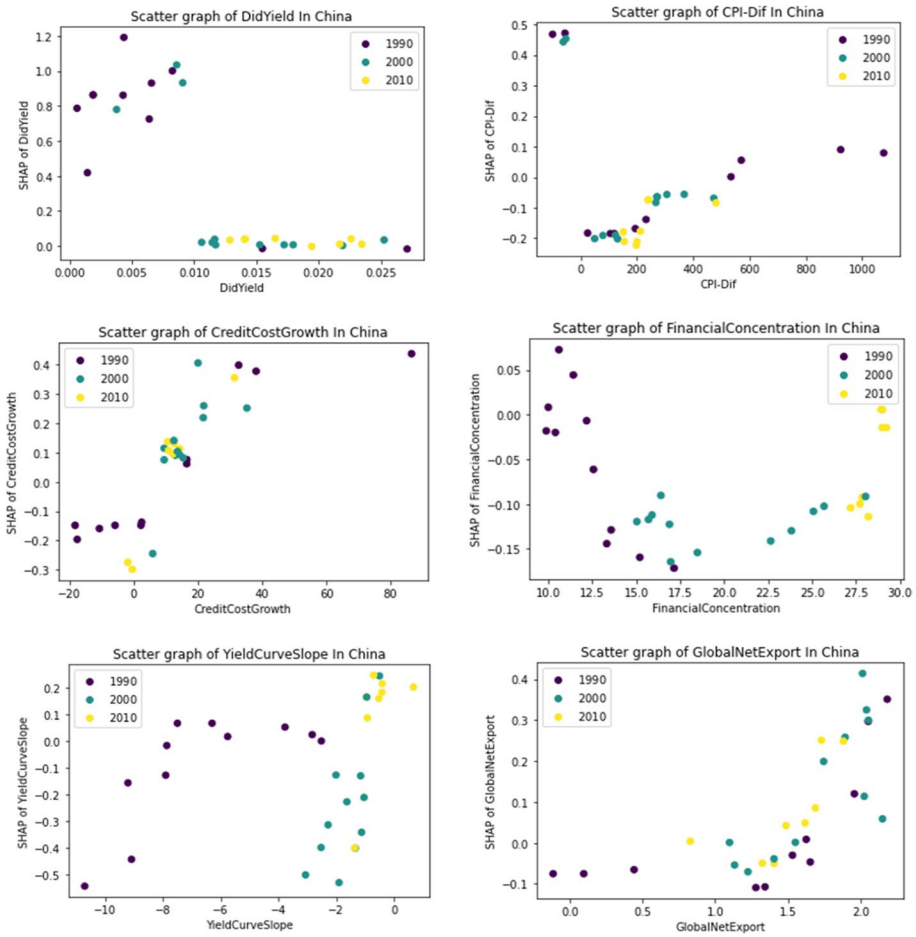


Fig. 7 Scatter plots showing the relationship between SHAP value and feature value in China data. Only a few are shown here. The others can be found in Additional file 2: Appendix S4. Figure legends are as follows: 1990 covers 1990–2000, and 2000 indicates 2000–2010, 2010 represents 2010–2020

consistent with the previous analysis. Since 2017, because of financial deleveraging and the US–China trade conflict, NarrowMoney, GlobalNetExport, and market expectations have been strong driving forces for risk.

Feature dimension

In the feature dimension, we drew scatter plots similar to the above but added time, as shown in Fig. 7, to examine the relationship between SHAP and the feature values in the China dataset.[22,23] Again, most feature plots exhibited trends similar to those in the global dataset.

Specifically, DidYield is a significant driving force for risk when DidYield is less than 0.01, and because the DidYield in China does not reach 0.07, a suitable threshold cannot be determined. Regarding time, China's DidYield has performed well most of the time since 2000, but there is still room for development.

[22] GBDT's plots with China data showed in Additional file 2: Appendix S6.
[23] China's PDP plots showed in Additional file 2: Appendix S7.

GlobalNetExport is in a position similar to DidYield. Because of the minimum limit, there is no pulling force on the left. However, on the right, the threshold in China, which is close to 1.5, is slightly smaller than the global threshold (2), possibly because China is export-oriented, which makes it more sensitive to the global trade environment.

In both the global and China datasets, YieldCurveSlope has the same threshold of -1, and when the feature value exceeds the threshold, it is more likely to occur at risk. In China, the yield curve slope has gradually increased since 1990, especially after 2010, most of it in the region of increasing risk. This indicates that the market has lowered economic expectations, and there is an obvious risk aversion, which is worthy of consideration.

Most NetExport values in China fall within the safe haven zone. The threshold and trend of CreditCostGrowth do not differ much from the global set, nor do the CPI-Dif trends.

For FinancialConcentration, whether in Global set or China set, its SHAP shows a tendency to fall first and then rise as FinancialConcentration grows. Numerically, when FinancialConcentration exceeds 13, it will reduce the probability of risk in most cases. While on the right, similar to GlobalNetExport and DidYield, the suitable threshold cannot be determined since the FinancialConcentration in China does not reach 35. It can also be seen from the combination of times that China's national financial concentration gradually increased from 1990 and performed well from 2000 to 2010, but it gradually approached the global alert boundary after 2010, which requires much attention.

NarrowMoney has always contributed to risk and shows a slight upward trend, which is consistent with the analysis above.

The pulling force of Credit is not obvious in China, which is inconsistent with the conclusions of most studies, possibly because the information in these two indicators is similar to that in CreditCostGrowth. Therefore, the model selects the factor with better performance and skips features with repeated information, indicating that credit cost is a better factor than credit itself for capturing risk.

Combined with the annual SHAP value trend in China above, it is evident that CreditCostGrowth is almost sensitive to each rise in risk. In addition, GlobalNetExport, NarrowMoney, and DidYield all float significantly during distinct risk increases, whereas DidYield slightly lags behind the others. These three indicators contributed to the risk calculation for the six periods. Taken together, the three indicators of CreditCostGrowth, NarrowMoney, and GlobalNetExport are more suitable for monitoring risk in China, not only because of the variation in the risk probability ratio caused by changes in the SHAP value, but also because of their high sensitivity to risk.

Shapley regression

The results of the Shapley regression on the Chinese dataset[24] are presented in Table 10. The regression coefficients in Table 10 reveal that most factors are statistically significant for measuring systemic financial risk in China, demonstrating the robustness of the indicators. For YieldCurveSlope and CreditCostGrowth, an increase of one unit in their SHAP values increases $\frac{y_{risk}}{1-y_{risk}}$ by 3.66% and 3.32%, respectively, which are the most influ-

[24] GBDT's table was in Additional file 2: Appendix S9.

Table 10 RF-Shapley regression with China data (sorted by the size of share)

Feature	Sign[a]	Share (SSCvalue)	Coefficient	*P* value	Exponentiation	Change of $\frac{y_{risk}}{1-y_{risk}}$
YieldCurveSlope	+	0.1660***	0.0359	0.000	1.0366	Significantly increased by 3.66%
Credit	+	0.1398***	0.0075	0.000	1.0075	Significantly increased by 0.75%
NarrowMoney	+	0.1382***	0.0116	0.000	1.0117	Significantly increased by 1.17%
Credit-Dif	+	0.1106***	0.0082	0.000	1.0082	Significantly increased by 0.82%
CreditCostGrowth	+	0.1091***	0.0327	0.000	1.0332	Significantly increased by 3.32%
DidYield	−	0.1081***	0.0311	0.000	1.0316	Significantly decreased by 3.16%
CPI-Dif	−	0.0823***	0.0190	0.000	1.0192	Significantly decreased by 1.92%
GlobalNetExport	+	0.0611***	0.0134	0.000	1.0135	Significantly increased by 1.35%
FinancialConcentration	−	0.0432***	0.0091	0.000	1.0091	Significantly decreased by 0.91%
NetExport	−	0.0417***	0.0062	0.000	1.0062	Significantly decreased by 0.62%

***, **and * respectively means that the null hypothesis is rejected at the significance level of 1%, 5%, 10%

[a] Since we don't have risk target of China, it is impossible to determine sign for China with Logistic Regression. Therefore, we took the feature value as the independent variable, the corresponding SHAP value as the dependent variable, and then fitted the line to obtain the sign. Moreover, to be robust, we averaged the SHAP value calculated by RF and GBDT for features.

ential factors for capturing risks in China. The increase in SHAP for Credit and Credit-Dif has a minor impact on risk, which we speculate is influenced by CreditCostGrowth, again suggesting that CreditCostGrowth is more appropriate than Credit and Credit-Dif for risk identification in China. A one-unit increase in DidYield's SHAP could decrease $\frac{y_{risk}}{1-y_{risk}}$ by 3.16%, which is different from the global dataset because from 1990 to 2020, DidYield in China has always been less than 0.01 and China's stock market is developing. In addition, consistent with a previous analysis in China, it is different from the training set in that a one-unit increase in NarrowMoney's SHAP could increase $\frac{y_{risk}}{1-y_{risk}}$ by 1.17%. In addition, an increase of one unit in NetExport's SHAP value decreases $\frac{y_{risk}}{1-y_{risk}}$ by 0.62%, which is consistent with China's export-oriented reality. Another difference is that GlobalNetExport increases risk, which is inconsistent with the above. Therefore, we draw another scatter[25] of GlobalNetExport in the global set after 1990. Most of the points fall on the right half of the U-shaped curve, which is understandable because there was more international trade and communication than before the 1990s. There is an upward trend, indicating that after 1990, GlobalNetExport's increase raised the risk. It is reasonable and similar to previous studies that international trade may increase risk contagion, and the larger the GlobalNetExport, the greater the likelihood of sovereign debt risk. Additionally, we note that the coefficient on CPI-Dif is also negative, which is inconsistent with the analysis, and we think it is because the effect of CPI-Dif on risk is non-linear and the regression is influenced by the deflation outlier because of the small

[25] The scatter can be found in Additional file 2: Appendix S10.

volume of Chinese data. For FinancialConcentration, according to the previous analysis, it has a concave impact on risk, and the feature value that is less than the minimum of the safe-haven zone has a greater effect on risk, so the overall coefficient appears to be negative.

Discussion and conclusion

This study upgrades a workflow consisting of selecting indicators, modeling, decomposing results, and statistical inference, enabling the workflow to accurately classify and interpret, as well as select the best subset of features. Compared with the old workflow, the new workflow has the advantage of selecting features and can lead to better model performance. In addition, the workflow in this study is well suited to categorical or regression studies in many areas. These techniques can also be applied independently or in combination, depending on the research requirements. Our results benefit from the distinct advantages of feature engineering techniques, including RFE, machine learning algorithms, RF and GBDT, cutting-edge interpretation models, SHAP values, and the Shapley regression statistical inference framework. The combination of these technological techniques can not only determine proper features and acquire high prediction precision, but also allow us to decompose the intrinsic functions of the model and conduct statistical inferences on all variables in non-linear, non-parametric models, combining the merits of difficult models and statistical inference. This study can serve as an example for others, and scholars can switch to other preferred shallow machine learning models or even deep learning algorithms for modeling.

We experimented with this in China's systemic financial risk research by investigating numerous impact factors. We collected training samples primarily from the Jordà–Schularick–Taylor Macrohistory Database, including 16 countries from 1870 to 2016, with a rich set of indicators, giving us an opportunity for feature engineering and providing us with many risk samples and properties. Inspired by international credit, we added four factors to the global trade environment to verify the usefulness of GlobalNetExport in predicting risk.

According to the selected samples and indicators, we captured six periods of rising risk probabilities in China since the 1990s, which are consistent with reality. We also detected the driving factors of each period, as well as the safe haven zone for each characteristic. In addition, we found that CreditCostGrowth, NarrowMoney, and GlobalNetExport are appropriate indicators for monitoring systemic risk in China, given their massive effect and high sensitivity to risk. More significantly, we confirmed the confidence of the variables and model. Finally, based on the analysis results, we offer the following recommendations:

First, for global economies, the cost of credit impacts risk more than credit itself. In addition, the prosperity of the stock market needs to be moderate because both excessively prosperous and underdeveloped stock markets have negative effects on systemic risk. In contrast, national financial concentration should not be too low because it has a great pull on risk through the channel of the government debt crisis. Additionally, international trade became more frequent after 1990. The scatter plots of the global trade environment during this time indicate that the higher the global average net export, the

greater its drive for systemic risk, signifying that countries should maintain a balanced current account to avoid debt accumulation, which could lead to a debt crisis and cause financial system turbulence.

Second, for countries with incomplete financial openness such as China, existing research demonstrates that global credit plays little role in the measurement of systemic risk, and this study shows that global trade is beneficial for China's risk calculation. Therefore, for similar countries, the global trade environment can also be considered when predicting risks or even in other research areas that need to consider the external environment.

There are also a few recommendations for China:

First, as noted above, the cost of credit is a better indicator of risk than the credit itself.

Second, Chinese stock market is not currently a major risk driver and there is room for development. China can prompt its stock market to take suitable measures to increase dividend rates, including guiding listed companies to develop reasonable dividend distribution plans.

Third, China needs to pay attention to the national financial concentration because it is increasing and moving closer to the border of the safe haven zone. Therefore, to avoid systemic risk, it is essential to optimize the tax structure, liberalize the domestic market, and achieve moderate financial concentration.

Fourth, China should focus on its international trade. Moreover, since the global trade environment will also have an impact on risks, and even its alignment with risks exceeds China's net export, it can be said that this is a new risk monitoring indicator, and China needs to pay close attention to global import and export exchanges and establish a dynamic management mechanism.

Fifth, the impact of narrow money on risk in China is distinctive: an increase in narrow money increases the crisis. The rise in NarrowMoney in China was driven mainly by the rapid increase in enterprise demand deposits and the increase in real estate demand deposits, which shows weak domestic demand in China. Therefore, China should expand domestic demand in a timely manner, which is also consistent with China's "dual circulation" development pattern.

Finally, because YieldCurveSlope is more likely to be prone to risk when it exceeds -1, and China's YieldCurveSlope has been rising and exceeded the threshold after 2010, the Chinese government should strengthen its communication with the market, guide positive market expectations, actively relieve risk aversion, and improve investor confidence.

The main limitation of this study was data availability, which is a typical drawback of macro studies. However, in future studies, Data mining methods can be used to collect data, find more distinguishing regulatory indicators that differ from macro features, and monitor systemic risk in a timely manner without suffering from data unavailability and delays. Further research is required to add submarket-specific features to the model. For example, the relationship between micro-characteristics and macro-indicators can be clarified using methods that deal with mixed-frequency data. Noneconomic factors, such as epidemics and politics, can also be added. In addition, other machine learning algorithms can be applied to predict systemic risk.

Abbreviations

RF	Random forest
GBDT	Gradient boosting decision tree
SHAP	Shapley additive explanations
RFE	Recursive feature elimination
MI	Mutual information
VIF	Variance inflation factor
CV	Cross-validation
ROC	Receiver operating characteristic
AUC	The area under curve
SSC	Shapley share coefficients
GDP	Gross domestic product
US	The United States
PDP	Partial dependence plot
PCA	Principal component analysis
IME	Interactions-based method for explanation
LASSO	Least absolute shrinkage and selection operator
FPR	False positive rate
KLR	Kaminsky–Lizondo–Reinhart
DCSD	Developing country studies division
CoVaR	Conditional value-at-risk
MSE	Mean squared error

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40854-023-00574-3.

Additional file 1: Spearman correlation matrix.

Additional file 2: Supplementary charts and descriptions.

Acknowledgements
Not applicable.

Author contributions
WD: The design of the study, supervision, reviewing and editing. ZYX: The design of the study, Methodology, writing original draft. All authors read and approved the final manuscript.

Funding
This work has been funded by National Social Science Fund of China (No. 22AGJ006).

Availability of data and materials
The datasets used during the current study are available from authors on reasonable request and the original data are available in the JORDÀ–SCHULARICK–TAYLOR MACROHISTORY DATABASE, http://www.macrohistory.net/data/ (Jordà et al. 2017, 2019).

Declarations

Competing interests
The authors declare no competing interests.

Received: 24 March 2022 Accepted: 5 December 2023
Published online: 11 March 2024

References

Abbritti M, Dell'Erba S, Moreno A, Sola S (2018) Global factors in the term structure of interest rates. Int J Cent Bank 14(2):301–340
Acharya V, Pedersen LH, Philippon T, Richardson M (2017) Measuring systemic risk. Rev Financ Stud 30(1):2–47. https://doi.org/10.1093/rfs/hhw088
Adrian T, Brunnermeier MK (2016) CoVaR. Am Econ Rev 106:1705
Altmann T, Bodensteiner J, Dankers C, Dassen T, Fritz N, Gruber S, Kopper P, Kronseder V, Wagner M, Renkl E (2020) Limitations of interpretable machine learning methods. Department of Statistics LMU Munich
Asgharian H, Hou AJ, Javed F (2013) The importance of the macroeconomic variables in forecasting stock return variance: a GARCH-MIDAS approach. J Forecast. https://doi.org/10.1002/for.2256
Ariza-Garzón MJ, Arroyo J, Caparrini A, Segovia-Vargas M (2020) Explainability of a machine learning granting scoring model in peer-to-peer lending. IEEE Access 8:64873–64890. https://doi.org/10.1109/ACCESS.2020.2984412
Baehrens D, Schroeter T (2010) How to explain individual classification decisions. J Mach Learn Res 11:1803–1831

Barbella D, Benzaid S, Christensen J, Jackson B, Qin XV, Musicant D (2009) Understanding support vector machine classifications via a recommender system-like approach. In: Proceedings of international conference on data mining DMIN, pp 305–311
Berg A, Pattillo C (1999) Predicting currency crises: the indicators approach and an alternative. J Int Money Financ 18(4):561–586. https://doi.org/10.1016/S0261-5606(99)00024-8
Bianchi F (2020) The great depression and the great recession: a view from financial markets. J Monet Econ 114:240–261. https://doi.org/10.1016/j.jmoneco.2019.03.010
Billio M, Pelizzon L, Savona R (2017) Systemic risk tomography: signals, measurement and transmission channels. Elsevier Ltd, Oxford
Bisias D, Flood M, Lo AW, Valavanis S (2012) A survey of systemic risk analytics. Office of Financial Research Working Paper No. 0001
Bluwstein K, Buckmann M, Joseph A, Kang M, Kapadia S, Şimşek Ö (2020) Credit growth, the yield curve and financial crisis prediction: evidence from a machine learning approach. Bank of England Staff Working Paper No. 848
Bracke P, Datta A, Jung C, Sen S (2019) Machine learning explainability in finance: an application to default risk analysis. Bank of England Working Paper No. 816
Breiman L (1996) Bagging predictors. Mach Learn 24:123–140. https://doi.org/10.1023/A:1018054314350
Breiman L (2001) Random forests. Mach Learn 1:15–32
Buckmann M, Joseph A (2022) An interpretable machine learning workflow with an application to economic forecasting. Bank of England Staff Working Paper No. 984
Buckmann M, Joseph A, Robertson H (2021) Opening the black box: machine learning interpretability and inference tools with an application to economic forecasting. In: Consoli S, Reforgiato Recupero D, Saisana M (eds) Data science for economics and finance. Springer, Cham. https://doi.org/10.1007/978-3-030-66891-4_3
Caggiano G, Calice P, Leonida L (2014) Early warning systems and systemic banking crises in low income countries: a multinomial logit approach. J Bank Finance 47:258–269. https://doi.org/10.1016/j.jbankfin.2014.07.002
Cardarelli R, Elekdag S, Lall S (2011) Financial stress and economic contractions. J Financ Stab 7(2):78–97. https://doi.org/10.1016/j.jfs.2010.01.005
Carmona P, Climent F, Momparler A (2019) Predicting failure in the U. S. banking sector: an extreme gradient boosting approach. Int Rev Econ Financ 61:304–323. https://doi.org/10.1016/j.iref.2018.03.008
Cesa-Bianchi A, Martin FE, Thwaites G (2019) Foreign booms, domestic busts: the global dimension of banking crises. J Financ Intermed 37:58–74. https://doi.org/10.1016/j.jfi.2018.07.001
Chakraborty C, Joseph A (2019) Machine learning at central banks. Bank of England Working Paper No. 674
Chan-Lau JA, Espinosa-Vega MA, Giesecke K, Sole JA (2009) Assessing the systemic implications of financial linkages. IMF Global Financial Stability Report Vol. 2
Chen SD, Wang Y (2014) Measuring systemic financial risk of China's financial institution——applying extremal quantile regression technology and CoVaR model. Chin J Manag Sci 22(7):10–17
Chen QL, Xue YC, Xiao L (2009) Early-warning of financial risk: indicator, mechanism, and empirical research. J Shanghai Univ (soc Sci Ed) 16(05):127–144
Chen H, Janizek JD, Lundberg S, Lee SI (2020) True to the model or true to the data? In: 2020 ICML workshop on human interpretability in machine learning, WHI 2020
Chu CCF, Chan DPK (2020) Feature selection using approximated high-order interaction components of the Shapley value for boosted tree classifier. IEEE Access 8:112742–112750. https://doi.org/10.1109/ACCESS.2020.3002665
Ecer F (2013) Comparing the bank failure prediction performance of neural networks and support vector machines: the Turkish case. Econ Res 3:81–98. https://doi.org/10.1016/j.eswa.2008.01.003
Ekinci A, Erdal HI (2017) Forecasting bank failure: base learners, ensembles and hybrid ensembles. Comput Econ 4:677–686. https://doi.org/10.1007/s10614-016-9623-y
Eygi B (2013) Prediction of bankruptcy using support vector machines: an application to bank bankruptcy. J Stat Comput Simul 83(8):1543–1555. https://doi.org/10.1080/00949655.2012.666550
Fan XY, Wang DP, Fang Y (2011) Measuring and supervising financial Instituites' marginal contribution to systemic risk in China: a research based on MES and leverage. Nankai Econ Stud 4:3–20
Fang Y (2016) Study on the transmission channel and measure of systemic risk: for macro-prudential policy implementation. Manag World 8: 32–57+187
Fang Y, Chen M, Yang YP (2018) The spillover effect and channel identification of financial market to banking systemic risk. Nankai Econ Stud 5:58–75
Fisher A, Rudin C, Dominici F (2018) All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. J Mach Learn Res 20(177): 1–81. http://arxiv.org/abs/1801.01489
Florez-Lopez R, Ramon-Jeronimo JM (2015) Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. A correlated-adjusted decision forest proposal. Expert Syst Appl 42(13):5737–5753. https://doi.org/10.1016/j.eswa.2015.02.042
Frankel JA, Rose AK (1996) Currency crashes in emerging markets: an empirical treatment. J Int Econ 41(3–4):351–366. https://doi.org/10.1016/S0022-1996(96)01441-9
Franck R, Schmied A (2003) Predicting currency crisis contagion from East Asia to Russia and Brazil: an artificial neural network approach. AMCB Working Paper
Frankel J, Saravelos G (2012) Can leading indicators assess country vulnerability? Evidence from the 2008–09 global financial crisis. J Int Econ 87(2):216–231
Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189–1232
Géron A (2017) Hands-on machine learning with Scikit-learn and TensorFlow, 1st edn. O'Reilly Media Inc, Sebastopol
Goldstein A, Kapelner A, Bleich J, Pitkin E (2015) Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. J Comput Graph Stat 24(1):44–65

Greenwood R, Landier A, Thesmar A (2015) Vulnerable banks. J Financ Econ 115(3):471–485. https://doi.org/10.1016/j.jfineco.2014.11.006
Guo N, Qi F, Zhang N (2018) Measurement and monitoring of systemic financial risk index in China. Finance Econ 02:1–14
Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Mach Learn 46(1):389–422
He DX, Yu JJ (2018) How to control the sluice gate of the money supply? Chin Rev Financ Stud 10(04): 1–10+118
Heaton J (2016) An empirical analysis of feature engineering for predictive modeling. SoutheastCon 2016:1–6. https://doi.org/10.1109/SECON.2016.7506650
Heiberger RH (2018) Predicting economic growth with stock networks. Physica A 489(1):102–111. https://doi.org/10.1016/j.physa.2017.07.022
Hilden J, Glasziou P (1996) Regret graphs, diagnostic uncertainty and Youden's Index. Stat Med 15(10):969–986. https://doi.org/10.1002/(SICI)1097-0258(19960530)15:10%3c969::AID-SIM211%3e3.0.CO;2-9
Huang ZG, Liu ZH, Zhu JL (2019) A general stack framework of credit risk rating models based on multi source data. J Quant Tech Econ 36(4):155–168
Illing M, Liu Y (2006) Measuring financial stress in a developed country: an application to Canada. J Financ Stab 2(3):243–265. https://doi.org/10.1016/j.jfs.2006.06.002
Jin T, Tao XY (2016) How government spending and opening up affect Chinese residents' consumption? On the basis of the exploration of Chinese transformational growth mode affecting consumption. China Econ Q 16(1):121–146
Johansson U, Konig R, Niklasson L (2004) The truth is in there—rule extraction from opaque models using genetic programming. In Proceedings of FLAIRS conference, pp 658–663
Jones S, Johnstone D, Wilson R (2015) An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes. J Bank Finance 56:72–85. https://doi.org/10.1016/j.jbankfin.2015.02.006
Jordà O, Schularick M, Taylor AM (2017) Macrofinancial history and the new business cycle facts. In: Eichenbaum M, Parker JA (eds) NBER macroeconomics annual 2016, vol 31. University of Chicago Press, Chicago
Jordà O, Knoll K, Kuvshinov D, Schularick M, Taylor AM (2019) The rate of return on everything, 1870–2015. Q J Econ 134(3):1225–1298. https://doi.org/10.1093/qje/qjz012
Joseph A (2019) Shapley regressions: a framework for statistical inference on machine learning models. Bank of England Staff Working Paper No. 784
Joseph A, Kalamara E, Kapetanios G, Potjagailo G (2021) Forecasting UK inflation bottom up. Bank of England Staff Working Paper No. 915
Joy M, Rusnák M, Šmídková K, Vašíček B (2017) Banking and currency crises: differential diagnostics for developed countries. Int J Financ Econ 22(1):44–67. https://doi.org/10.1002/ijfe.1570
Kaminsky G (1999) Currency and banking crises: the early warning of distress. IMF Working Paper No. 1999/178, 1999
Kaminsky G, Lizondo S, Reinhart C (1997) Leading indicators of currency crises. IMF staff paper, vol 45, no 1
Khalid S, Khalil T, Nasreen S (2014) A survey of feature selection and feature extraction techniques in machine learning. In: 2014 Science and information conference, pp 372–378. https://doi.org/10.1109/SAI.2014.6918213
Kinkyo T (2019) A bi-annual forecasting model of currency crises. Appl Econ Lett 27(4):255–261
Kou G, Xu Y, Peng Y, Shen F, Chen Y, Chang K, Kou SM (2021) Bankruptcy prediction for SMEs using transactional data and two-stage multiobjective feature selection. Decis Support Syst 140:113429. https://doi.org/10.1016/j.dss.2020.113429
Kou G, Chao X, Peng Y, Alsaadi FE, Herrera-Viedma E (2019) Machine learning methods for systemic risk analysis in financial sectors. Technol Econ Dev Econ 25(5):716–742
Kumar IE, Venkatasubramanian S, Scheidegger C, Friedler S (2020) Problems with Shapley-value-based explanations as feature importance measures. arXiv:2002.11097
Laeven L, Valencia F (2013) Systemic banking crises database. IMF Econ Rev 2:225–270
Lainà P, Nyholm J, Sarlin P (2015) Leading indicators of systemic banking crises: Finland in a panel of EU countries. Rev Financ Econ 24(January):18–35
Li MB, Liang S (2021) Monitoring systemic financial risks: construction and state identification of China's financial market stress index. J Financ Res 06:21–38
Li T, Kou G, Peng Y, Yu PS (2022) An integrated cluster detection, optimization, and interpretation approach for financial data. IEEE Trans Cybern 52(12):13848–13861. https://doi.org/10.1109/TCYB.2021.3109066
Lo DM, Peltonen TA (2013) Assessing systemic risks and predicting systemic events. J Bank Finance 37(7):2183–2195. https://doi.org/10.1016/j.jbankfin.2012.06.010
Locklin S (2014) Neglected machine learning idea. Neglected machine learning ideas | Locklin on science (wordpress.com). Accessed 28 Oct 2021
Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: NIPS'17: proceedings of the 31st international conference on neural information processing, pp 4768–4777
Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, Liston DE, Low DK, Newman S, Kim J, Lee S (2018) Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng 2:749–760. https://doi.org/10.1038/s41551-018-0304-0
Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal B, Lee SI (2020) From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2:56–67. https://doi.org/10.1038/s42256-019-0138-9
Ma Y (2013) Monetary policy framework based on financial stability: theoretical and empirical analysis. Stud Int Finance 11:4–15
Mao R, Liu NN, Liu R (2018) The expansion of local government debt and the mechanism of systemic financial risk triggering. China Ind Econ 04:19–38
Molnar C (2023) Interpretable machine learning: a guide for making black box models explainable, 2nd edn. christophm.github.io/interpretable-ml-book/

Momparler A, Carmonab P, Climent F (2016) Banking failure prediction: a boosting classification tree approach. Span J Finance Account 1:63–91. https://doi.org/10.1080/02102412.2015.1118903
Motoda H, Liu H (2002) Feature selection, extraction and construction. Commun Inst Inf Comput Mach 5:67–72
Nag A, Mitra A (1999) Neural networks and early warning indicators of currency crisis. Reserve Bank of India Occasional Papers, vol. 20, no.2
Nivorozhkin E, Chondrogiannis I (2020) Shifting balances of systemic risk in the Chinese banking sector: Determinants and trends. J Int Financ Mark Inst Money 76:101465. https://doi.org/10.1016/j.intfin.2021.101465
Reinhart CM, Rogoff KS (2008) Is the 2007 US sub-prime financial crisis so different? An international historical comparison. Am Econ Rev 98(2):339–344. https://doi.org/10.1257/aer.98.2.339
Research Team at center for finance and development, Tsinghua National Institute of Financial Research (2019) Research on monitoring systemic financial stress in China. Stud Int Finance 12:3–12
Ribeiro MT, Singh S, Guestrin C (2016) "Why should I trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 1135–1144. https://doi.org/10.1145/2939672.2939778
Roy S (2009) Predicting the Asian currency crises with artificial neural networks: what role of function AP proximation? Economics (2)
Sachs J, Tornell A, Velasco A (1996) The Mexican peso crisis: sudden death or death foretold? J Int Econ 41(3–4):265–283. https://doi.org/10.1016/S0022-1996(96)01437-7
Schularick M, Taylor AM (2012) Credit booms gone bust: monetary policy, leverage cycles, and financial crises, 1870–2008. Am Econ Rev 102(2):1029–1061. https://doi.org/10.3386/w15512
Savage N (2022) Breaking into the black box of artificial intelligence. Nature. https://doi.org/10.1038/d41586-022-00858-1
Sekmen F, Kurkcu M (2014) An Early warning system for Turkey: the forecasting of economic crisis by using the artificial neural networks. Asian Econ Financ Rev 4(4):529
Shapley LS (1953) A value for n-person games. In: Kuhn HW, Tucker AW (eds) Contributions to the theory of games. Princeton University Press, Princeton, pp 307–317
Shi JP, Gao Y (2009) A study on KLR financial crises warning model. J Quant Tech Econ 26(3):106–117
Shrikumar A, Greenside P, Kundaje A (2017) Learning important features through propagating activation differences. arXiv:1704.02685
Son H, Hyun C, Phan D, Hwang HJ (2019) Data analytic approach for bankruptcy prediction. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2019.07.033
Strumbelj E, Kononenko I (2010) An efficient explanation of individual classifications using game theory. J Mach Learn Res 11:1–18
Su DW, Xiao ZX (2011) Research on early warning system for financial crisis in China: evidence based on macroeconomic data from six Asian countries. Stud Int Finance 6:14–24
Suss J, Treitel H (2019) Predicting bank distress in the UK with machine learning. Bank of England Staff Working Paper No. 831
Tao L, Zhu Y (2016) On China's financial systemic risks. J Financ Res 432(6):8–36
Tölö E (2020) Predicting systemic financial crises with recurrent neural networks. J Financ Stab 49:100746. https://doi.org/10.1016/j.jfs.2020.100746
Wang KD (2019) The selection of early-warning models and forward-looking indicators of financial crisis. Financ Regul Res 8:84–100
Wang Q, Tian J (2016) Banking capital supervision and systemic financial risk transfer—an analysis based on the DSGE model. Soc Sci China (03):99–122+206–207.
Wang D, Zhou YX (2020) Application of random forest model in macro prudential regulation—an empirical study based on the data of 18 countries. Stud Int Finance 11:45–54
Wang T, Yuan CG, Wang CY (2020) Does applying deep learning in financial sentiment analysis lead to better classification performance? Econ Bull 40(2):1091–1105
Ward F (2017) Spotting the danger zone: forecasting financial crises with classification tree ensembles and many predictors. J Appl Economet 32(2):359–378. https://doi.org/10.1002/jae.2525
Wei JN, Yang GP (2017) Shocks of two international financial crises on China and its countermeasures. New Finance Rev 01:104–145
Wu CW, Liang JH, Wang W, Li CS (2017) Random forest algorithm based on recursive feature elimination. Stat Decis 21:60–63
Xiong C, Jin H (2018) Double helix of local government debt risk and financial sector risk—analysis based on nonlinear DSGE model. China Ind Econ 12:23–41
Yan DW, Chi GT, Lai KK (2020) Financial distress prediction and feature selection in multiple periods by lassoing unconstrained distributed lag non-linear models. Mathematics 8:1275. https://doi.org/10.3390/math8081275
Yang ZH, Wang SD (2021) Systemic financial risk contagion of global stock market under public health emergency: empirical evidence from COVID19 epidemic. Econ Res J 56(08):22–38
Yang ZH, Chen YT, Xie RK (2018) Research on systemic risk measures and cross-sector risk spillover effect of financial institutions in China. J Financ Res 460(10):19–37
Yang ZH, Chen YT, Chen LX (2019) Effective measurement and nonlinear contagion of extreme financial risk. Econ Res J 54(5):63–80
Yang ZH, Chen YT and Zhang PM (2020) Macroeconomic shock, financial risk transmission and governance response to major public emergencies. J Manag World 36(05): 13–35+7. https://doi.org/10.19744/j.cnki.11-1235/f.2020.0067
Zaidi NA (2015) Feature engineering in machine learning. https://doi.org/10.13140/RG.2.1.3564.3367
Zhang AJ (2015) Comparative analysis of dynamic warning of National Financial Security (1992–2011). World Econ Stud (04):3–12+127. https://doi.org/10.13516/j.cnki.wes.2015.04.002
Zheng A, Casari A (2018) Feature engineering for machine learning, 1st edn. O'Reilly Media Inc, Sebastopol
Zhou H, Wang Q (2009) Monetary policy and asset price volatility: theoretical model and empirical study in China. Econ Res J 44(10):61–74

Zhou KG, Xing ZY, Peng SY (2020) The contagion mechanism between industrial risk and the macro economy in China. J Financ Res 12:151–168
Zięba M, Tomczak SK, Tomczak JM (2016) Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. Expert Syst Appl 58:93–101. https://doi.org/10.1016/j.eswa.2016.04.001

Publisher's Note