ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
https://doi.org/10.17559/TV-20231021001043
Original scientific paper

An XGboost Algorithm Based Model for Financial Risk Prediction

Yunsong XU, Jiaqi LI*, Anqi WU

Abstract: This study presents a novel financial risk prediction model utilizing the XGboost algorithm, analyzing macroeconomic data from the Jorda-Schularic-Taylor database. Our method achieves an 84.77% *accuracy* rate in predicting systemic financial risks. Unlike traditional models, this model combines the anomaly detection algorithm with the XGboost model, solving the possible "gray sample" problem and improving predictive accuracy. The model's feature importance analysis reveals key indicators, providing insights into the dynamics of financial risk occurrence. Finally, the systemic financial risk score is used to comprehensively evaluate a country's systemic financial risk level, offering a robust risk assessment and monitoring tool. This research enhances the application of machine learning in financial risk prediction, offering a reference for improving risk identification and prevention.

Keywords: machine learning; prediction model; systemic financial risk; XGBoost

1 INTRODUCTION

The subprime mortgage crisis sweeping the globe in 2008 dramatically increased regulatory authorities' attention to systemic financial risks. The Financial Stability Board (FSB) pointed out in 2019 that underestimating systemic financial risks or overestimating financial system resilience could trigger a global financial crisis, resulting in severe economic, social, and political consequences. Therefore, preventing and resolving systemic financial risks is a topic that the financial sector and even the entire society cannot avoid.

A forward-looking forecast of systemic risks is essential for preventing and resolving systemic financial risks, and many scholars have worked in this area. Early scholars favored logit regression approaches, which they viewed as less biased and more accurate in estimating the probability of systemic financial risk occurrence [1, 2]. However, the estimation based on linear approaches is gradually considered unable to predict systemic financial risks accurately. With the development of computer technology, machine learning techniques and models, such as random forests, neural networks, and support vector machines, have been applied to study systemic financial risks [3, 4].

At present, the application of machine learning to predict systemic financial risks has made significant progress, but there are still some problems to be solved. First, supervised learning models need to be applied correctly. The occurrence of systemic financial risks is a low-probability event, so the sample data is almost certainly highly unbalanced, which can easily lead to bias in model training if the data is not handled correctly. Second, there is no exact criterion for judging whether systemic financial risks occur, so the data scholars use are subjective to some degree. In other words, there may be cases where systemic financial risks should be identified but are not identified, also known as "gray data" and can easily result in selection bias [5]. Third, the interpretability of the model is poor [6]. In fact, a "black box" type of forecasting model cannot meet economic forecasting requirements. Fourth, the overfitting problem of machine learning should be emphasized [7]. Fifth, due to the complexity of systemic financial risks, nonlinear and higher-order data characteristics should be applied more frequently. Based on the above problems, this article constructs a systematic financial risk prediction system, combining the anomaly detection algorithm and the XGboost algorithm, and discusses the model's interpretability in detail. It makes up for the defects of insufficient interpretability of existing research and inadequate attention to the gray sample, improving the model's accuracy. It is significant for strengthening the application of machine learning in systemic financial risk early warning.

2 LITERATURE REVIEW

Early systemic financial risk prediction models mainly adopted the "signal extraction" technique. That is, to extract some abnormal signals from macroeconomic variables, such as financial and fiscal variables, to predict the occurrence of future financial crises and then build an early warning system. The studies of Kaminsky and Reinhart and Christensen and Li are representative, the former using indicator systems and the latter using comprehensive indicators [8, 9].

However, Davis and Karin question the signal extraction technique [10]. They argue that this technique does not reflect the differences among countries and is, therefore, unsuitable for early warning on a global scale. In contrast, the predictive ability of logit regression is superior, using which scholars have identified many variables with good predictive effects on systemic financial risks, such as capital adequacy ratio, liquidity ratio, and the growth rate of real estate prices; the ratio of credit to GDP [1, 11]. Filippopoulou et al. also used the logit model to assess the predictive ability of the risk indicators from the European macroprudential database, and the results showed that these indicators could anticipate crises effectively [12].

The logit model essentially performs a classification task. With the abundance of data and the advancement of computer technology, many machine learning models show excellent performance in this regard, such as the traditional methods of support vector machines, extreme random trees, stochastic gradient descent, and artificial neural networks, as well as methods that developed based on them, such as bi-directional long-short term memory models, support vector machine rule extraction, and neural networks based on entropy theory [13-15]. Due to their excellent performance, they are also applied in micro-level business scenarios, such as customer experience prediction, online loan credit risk measurement, and employee job satisfaction prediction [16-19]. In recent years, machine learning models have come into the view of systemic financial risk researchers and provide a novel perspective for research.

The decision tree model is a straightforward machine learning model and was earlier applied to predict systemic financial risk. It can be divided into two types based on the output-classification trees for discrete output and regression trees for continuous output. Manasse and Roubini (2009) studied sovereign debt crises using the above two decision tree types and proposed a rule of thumb to identify government default characteristics under different crisis causes [20]. Savona and Vezzoli (2015) constructed an early warning model for sovereign debt crises using the improved regression tree and regarded illiquidity, historical default rates, real GDP growth, and interest rates on U.S. debt as determinants of sovereign debt crises in emerging market countries and Europe [21].

With the development of technology, the ensemble learning model was applied more often. This model combines single models, such as the decision tree model, to improve model performance. As one of the relatively mature models, the random forest model based on the decision tree model has been applied earlier to predict systemic financial risk. Alessi and Detken (2018) used the model to predict banking crises in EU countries during 1970 - 2012 [22]. Joy et al. (2017) used the model to predict banking and currency crises in 36 developed economies [23].

The Adaptive Boosting (Adaboost) algorithm, also based on the decision tree model, pays more attention to the misclassified data during training. This algorithm boosts the weight of misclassified samples during iteration to learn more from the misclassified samples and improve model performance. Casabianca et al. (2019) used the Adaboost model to forecast the financial risks of 100 developed economies and emerging economies [24]. Fouliard et al. (2020) constructed a new financial risk prediction model by combining the decision tree-based and regression models [25]. These models, without exception, indicate that the predictive power of machine learning methods for financial risks is significantly better than that of traditional logistic regression.

In recent years, more new machine learning models have been applied to systematic financial risk prediction, among which the human worker neural network and the support vector machine model are representative. The basic idea of the support vector machine model is to classify by finding the optimal decision boundary, which has good generalization performance and high accuracy. SVM has also been applied to financial risk prediction for a long time, but it was primarily used in microfields in the early days. For example, Kim (2003) studied the predictive power of SVM on stock indexes, and Lee (2006) used the model to predict enterprise credit rating [26, 27]. Ahn et al. (2010) used SVM to construct a financial crisis early warning system and found it was effective in predicting financial crises [28].

In terms of the artificial neural network, Ristolainen constructed a systemic financial risk early warning system using an artificial neural network model, and he pointed out that the artificial neural network predicted much better than traditional linear approaches represented by the logit model due to the differences among countries [3]. Tölö constructed a systematic financial risk prediction model using a recurrent neural network and argued that the recurrent neural network could provide more robust and consistent prediction results [29]. Bluwstein compared some machine learning models and their advantages over logistic regression, proving they were more suitable for systemic financial risk early warning [4].

However, all the above models have room for improvement. Ristolainen and Bluwstein evaluated the performance of machine learning models but did not discuss their interpretability [3, 4]. Tölö discussed the model interpretability but selected too few feature variables, just five, which failed to exert machine learning models' capabilities [29]. Even essentially, we might not need a pretty complex model if we can a priori select the features in a very small range. Another problem with Tölö is that the database contained too few positive samples (the samples having experienced systemic financial risks) [29]. When we fail to address the problem correctly, it will lead to bias and potentially cause the model test set not to be allocated or allocated quite few positive samples. At this point, although we can evaluate models using *accuracy*, *AUC*, and other indicators, it is difficult to use indicators such as *precision* and *recall* rate, while this article holds that the *recall* rate is highly critical to the issue of systemic financial risks.

According to these problems, this article, taking the XGBoost algorithm as the basis and combining it with anomaly detection algorithms, constructs a systematic financial risk prediction model and discusses the model interpretability using the *SHAP* method in an attempt to provide references for further deepening the application of machine learning technology in the field of systematic financial risk prediction.

3 MODEL PRINCIPAL INTRODUCTION

The XGBoost algorithm is the primary method used in this article to construct the prediction model, whose full name is Extreme Gradient Boosting, and was first proposed by Chen and Guestrin [30]. XGBoost belongs to the gradient boosting decision tree (GBDT) algorithm but is superior to traditional GBDT algorithms in terms of algorithm *precision*, speed, and generalization ability. Regarding algorithm *precision*, XGBoost expands the loss function to the second-order derivative, bringing it closer to the actual loss. Regarding algorithm speed, weighted quantile sketch and sparse perception algorithm techniques are used to improve algorithm speed through cache optimization and model parallelism. Regarding algorithm generalization ability, model overfitting is prevented by adding regularization terms to the loss function, setting reduction rates in additive models, and column subsampling.

Specifically, assuming that each sample consists of a vector x_i composed of some feature indicators, and an output indicator y_i, letting sample set $S = \{(x_i, y_i)\}$,

$\left(i=1,2,3,\ldots,n,x_i \in R^m, y \in R\right)$, where m is the feature dimension, and n is the number of samples, the loss function of the model can be expressed as:

$$loss^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{h=1}^{t} \Omega\left(f_h\right) \quad (1)$$

where $\hat{y}_i$ is the model predicted value and y_i is the corresponding true value; $l(,)$ is the empirical *loss* function, and the XGBoost algorithm requires the loss function to be second-order derivable; f_h is the base model at the *th* iteration, and the XGBoost model usually chooses the CART classification tree as the base model; At this point, $\sum_{h=1}^{t} \Omega\left(f_h\right)$ is a regularization term representing the sum of the complexity of all base models and used by the XGBoost model to control the overfitting problem.

According to the forward stagewise algorithm, the predicted value in the model at step t can be expressed as:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t\left(x_i\right) \quad (2)$$

where $\hat{y}_i^{(t-1)}$ is the predicted value given by the model at step $t-1$, and exists as a known constant at step t. Likewise, the complexity of base models at the first $t-1$ steps has been determined so the regularization term can be split:

$$\sum_{h=1}^{t} \Omega\left(f_h\right) = \Omega\left(f_h\right) + \text{Constant} \quad (3)$$

Constant represents the sum of the complexity of the first $t-1$ base models, a constant here. Therefore, the loss function can be rewritten as:

$$\text{loss}^{(t)} = \sum_{i=1}^{n} \text{l}(\text{y}_i, \hat{\text{y}}_i^{(t-1)} + \text{f}_t(\text{x}_i)) + \Omega(\text{f}_t) + \text{Constant} \quad (4)$$

Performing a second-order Taylor expansion on the summation part and removing the constant term yield:

$$l\left(y_i, \hat{y}_i^{(t-1)}\right) + f_t\left(x_i\right) = l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f\left(x_i\right) + \frac{1}{2} h_i f_t^2\left(x_i\right) + \Omega\left(f_t\right) \quad (5)$$

where g_i is the first-order derivative of the loss function, and h_i is the second-order derivative of the loss function. As long as we ask for the first-order and second-order derivatives of the loss function at each step and then optimize the solution, we can get the $f(x)$ for each step of the model in the forward stagewise and thus obtain the XGBoost model by model additivity.

4 MODEL TRAINING AND EVALUATING
4.1 Data Sources

All the data used come from the Jorda-Schularic-Taylor macroeconomic database, which records macroeconomic data for 18 developed economies, including the United States, the United Kingdom, Belgium, and Denmark, from 1870 to 2020, and its long-time span and rich data types have made it widely used in academic research. In particular, the database records the time when systemic financial risks occurred in each country and represents it as a dummy variable, which makes it possible to construct supervised learning models using the XGBoost algorithm. One problem, however, is that the database indicates that no country in the sample should be recognized as having experienced systemic financial risks since 2008. Therefore, the dataset for this article is selected from 1870 to 2008, and a total of 2452 samples are obtained after excluding the data with missing systemic financial risk dummy variables.

It should be noted that the Jorda-Schularic-Taylor dataset also has certain limitations. First, despite covering an extended period, it cannot provide complete macroeconomic indicators for all countries. Especially for some historical data, there may be problems with missing or incoherent data, affecting study breadth and depth.

Second, the Jorda-Schularic-Taylor dataset covers a limited number of countries, and all of them are developed economies in Europe and the United States, which may bias the results as it prevents us from capturing characteristics of other developing countries or emerging economies.

4.2 Feature Variable Selection

According to the European Systemic Risk Board (ESRB), systemic risk indicators can be classified into six categories: credit development, private sector debt stress, potential asset overvaluation, external imbalance, risk mispricing, and bank balance sheet robustness. Tölö summarized and verified the systemic risk indicators proposed by previous research in these six aspects [31].

Table 1 Systemic financial risk feature variables

Variable type	Variable name	Content
Monetary credit level	gl/gdp	Annual growth rate of total non-financial sector loans to GDP
	Money	Nominal monetary volume
Private sector debt pressure	tloan	Total non-financial sector loans
	tmort	Non-financial sector mortgages
	bdebt	Corporate debts
	thh	Total household loans
External imbalance	cagdp	Current account balance divided by GDP
Potential overvaluation of assets	hpnom	Housing prices
	housing_tr	Gross housing yield
	eq_tr	Total return on stock
balance sheet of banks	lev	Bank capital adequacy ratio
	ltd	Bank deposit-to-loan ratio
	noncore	Bank non-core capital adequacy ratio
Mispricing of risks	stir	Short-term interest rate
	ltir	Long-term interest rate
Government financial situation	debtgdp	the ratio of public debt to GDP
	revenue	Government revenue
	expenditure	Government expenditure
	bond_tr	Government bond yield
Macroeconomic operation	rgdp_growth	Real per capita GDP growth rate
	unemp	Unemployment rate
	Iy	the ratio of investment to GDP

Tölö further pointed out that the five variables of private sector loans/GDP, the annual growth rate of real stock prices, the annual growth rate of housing prices, the annual growth rate of real GDP, and current account/GDP are the features most closely associated with the occurrence of systemic financial risks, and constructed a recurrent neural network model with these five feature variables to predict systemic financial risks [29].

Based on the conclusions of Tölö and Tölö, combined with existing research, this article finally selected 22 indicators in eight categories as feature variables, all of which lagged by one period. The specific variables are shown in the following table [29, 31]. Missing values have been filled with the mean values of respective features.

4.3 Data Set Construction

4.3.1 Anomaly Detection

The data used starts at a relatively old date, so there may be distortions. Meanwhile, there is no accepted objective criterion for judging whether systemic financial risks occur, so there may be a possibility that systemic financial risks have happened or should be recognized as having occurred but are not recorded, and the existence of such a "gray sample" will make the model training results doubtful [32]. Therefore, this article first adopts anomaly detection algorithms to eliminate anomalous data from the data not experiencing systematic financial risks to improve subsequent training accuracy.

In this article, two anomaly detection algorithms are selected. One is the Local Outlier Factor (*LOF*) algorithm proposed by Breuning et al., which puts forward the concept of "local reachability density" and determines whether a point is an outlier by calculating the relative density of the point and its neighboring points [33]. Specifically, it is assumed that the K distance of a certain data point P, the distance between the K-th nearest point to point P and point P, is $P_{k\text{-distance}}$. The distance here can be measured in different ways, and Min-distance is used in this article. The circle with point P as the center and $P_{k\text{-distance}}$ as the radius is called the k-distance neighborhood of point P. The maximum value between the direct distance between point P and another point O and the k distance of point P is referred to as the "reachability distance" of point P and point O. Then, the reciprocal of the average reachability distance between point P and the points in the k-distance neighborhood of point P is called the local reachability density. Obviously, the greater the local reachability distance, the smaller the density and the more "sparse." Local reachability density can be expressed as:

$$lrd_k(P)=\frac{1}{\frac{\sum_{O\in N_{k(p)}} reach\_disk_k(p,o)}{\left|N_{k(p)}\right|}} \tag{6}$$

where $N_{(k(p))}$ is the number of points in the k distance neighborhood of point P, and *reach_dist* is the reachability distance between point P and point O. At this point, the Local Outlier Factor (*LOF*) can be defined as the ratio of the k adjacent average local reachability density of point P to its local reachability density, that is:

$$LOF=\frac{\frac{\sum_{O\in N_{k(p)}} lrd(po)}{\left|N_{k(p)}\right|}}{lrd(p)} \tag{7}$$

It can be seen that the smaller the local reachability density of point p relative to its k adjacent average local reachability density, the more "outlier" the sample, the greater the outlier factor. In this paper, the number of neighbor points k is set to 25, and the threshold of neighborhood density is set to 0.7, a conservative value. This is because systemic financial risks generally have an "intense" manifestation, and there should not be many cases where risks occur but are not recorded. Therefore, such a setting reduces the misclassification of normal data points as anomalies. The distance between points is calculated using Euclidean distance.

A total of 484 outliers are detected based on the anomaly detection algorithm *LOF*, and the non-systematic financial risk samples among them are removed from the training set:

The second method used for anomaly detection is the Isolation Forest (IF), an unsupervised anomaly detection method that was first proposed by Liu et al. [34]. The algorithm identifies outliers in the dataset by constructing multiple decision trees. Compared with traditional anomaly detection methods, its linear time complexity and high accuracy make it ideal for big data processing scenarios. Its basic principle is as follows:

(1) For each data point, the algorithm constructs a decision tree with a randomly selected dimension as the root node.

(2) For other data points, the algorithm descends along the path of the decision tree until it reaches a leaf node.

(3) At leaf nodes, the algorithm compares the data point's value with the node's threshold. If the data point's value is less than the threshold, the data point is normal; otherwise, the data point is considered an outlier.

(4) The algorithm can obtain a final anomaly detection result by averaging multiple decision trees.

The isolated forest method is used for anomaly detection, and the model parameters are as follows:

Table 2 Model parameters

Parameter	Name	Value
n_estimators	Number of decision trees in the forest	100
min_samples_split	Minimum number of samples for tree splits	2
min_samples_leaf	Minimum number of samples for leaf nodes	1
contamination	Expected proportion of outliers	0.05

According to the anomaly detection result, 54 outliers are eliminated. The distribution of outliers is shown in the Fig. 1.

After removing outliers, the Adaptive Synthetic Sampling (ADASYN) oversampling method is used to process the non-outlier samples. It is an improvement on the Synthetic Minority Oversampling Technique (SMOTE). Specifically, ADASYN compensates for SMOTE's inability to handle distribution differences among categories well and can decide how many synthetic samples need to be generated for each minority sample through algorithms instead of generating the same data as in the case of SMOTE. Following the ADASYN oversampling processing, 2261 training samples are obtained, with a ratio of 1:1 between samples that experienced financial crisis and those that did not.

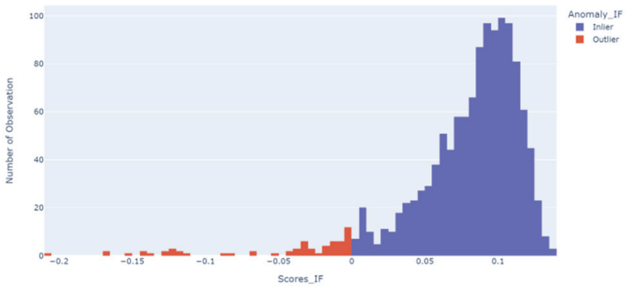

Figure 1 Outlier distribution based on the IF algorithm

4.4 Model Parameter Selection

In the XGBoost algorithm, the model parameter selection is closely related to the final training results. In this article, the Grid search cross-validation method is used for parameter adjustment, and considering the need to control model overfitting, the main model parameters are determined as follows:

Table 3 Parameter determination of the XGBoost model

Parameter	Name	Value
learning_rate	Learning rate	0.01
max_depth	Maximum depth of subtree	9
n_estimators	Maximum number of iterations	981
min_child_weight	Sub-node weight threshold	1
subsample	Training Sample Sampling Ratio	0.7
reg_lamda	L2 regularization coefficient	600

4.5 Model Training Results

4.5.1 Model Evaluation

The sample set is divided into a training set and a test set in a 7:3 ratio to train the model. Then, an evaluation index system is adopted to assess the model performance. For classification tasks, the commonly used evaluation index system mainly includes the following:

Accuracy, it means the proportion of samples predicted to be positive in all samples. If *TP*, *TN*, *FP*, and *FN* are used to represent true positive, true negative, false positive, and false negative, the formula for *accuracy* rate can be expressed as:

$$Accurancy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

Precision, it means the proportion of correctly predicted samples in all samples predicted to be positive. This indicator primarily compensates for the *accuracy* error caused by an unbalanced sample distribution. For example, when the positive proportion is too high, the prediction inclined to positive always has a higher *accuracy* rate. The *precision* rate formula can be expressed as:

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

Recall, it means the proportion of samples predicted to be positive in true positive samples; that is, the proportion of true positive samples that are correctly predicted. *Recall* reflects the degree to which we can recognize positive rather than focusing on overall accuracy. In our issue, *recall* plays a significant role because systemic financial risks are so destructive to an economy that even if our model makes a wrong prediction that systemic financial risks will occur, the resulting economic and social costs are usually smaller than those caused by incorrectly predicting that no systemic financial risks will occur. The *recall* rate formula can be expressed as:

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

The *F*1 score considers both *recall* and *precision* and is calculated as follows:

$$F1 = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision} \tag{11}$$

The indicators of the model in this article are shown in the table below:

Table 4 Model evaluation index

Model	*Accuracy*	*Precision*	*Recall*	*F*1	*AUC*
XGBoost	0.8477	0.8559	0.8523	0.8541	0.9149
Logistic	0.6644	0.8559	0.8523	0.8541	0.6522

As can be seen, the model shows good predictive ability, especially the *recall* rate, the main focus of this article, reaches 85.23%; in other words, the model can anticipate systemic financial risks with an *accuracy* of 85.23%. The same task is carried out with the traditional logistic regression method as a control. It is discovered that the *AUC* and *accuracy* rate of the model are significantly lower than those of the XGBoost algorithm, and its *precision* and *recall* rate are identical to those of the XGBoost algorithm mainly because the test set contains few positive samples. As the occurrence of systemic financial risks is a complex process, it may be difficult to characterize the mechanism using linear relationships, and the higher-order and nonlinear information provided by XGBoost contributes to constructing better prediction models:

Stability of the model over time: To verify the stability of the algorithm over time, this paper adopts a rolling modeling approach, taking 70 years as a time window, rolling forward ten years each time, and training the model separately according to the above process. Based on this approach, models are formed for eight time periods: 1870 - 1940, 1880 - 1950, 1890 - 1960, 1900 - 1970, 1910 - 1980, 1920 - 1990, 1930 - 2000, and 1940 - 2008. The *accuracy*, *recall*, *precision*, *F*1 score, and *AUC* value of the models for each period are shown in the table (the eight-time windows denoted by 1 - 8, respectively):

It can be seen that the model consistently performs well with the passage of time window, indicating that the systemic financial risk prediction system constructed in this paper has time stability.

Table 5 Stability of the model over time

Time	*Accuracy*	*Precision*	*Recall*	*F*1	*AUC*
1	*0.8899*	*0.8609*	*0.9252*	*0.8919*	*0.9586*
2	0.7885	0.7246	0.9091	0.8065	0.8736
3	0.8803	0.8507	0.9344	0.8906	0.9647
4	0.8697	0.8346	0.9250	0.8775	0.9475
5	0.9091	0.8794	0.9688	0.9219	0.9566

6	0.9559	0.9417	0.9741	0.9576	0.9839
7	0.9440	0.9370	0.9597	0.9482	0.9827
8	0.9237	0.8980	0.9706	0.9329	0.9755

4.5.2 Feature Contribution Analysis of the Model

In addition to the predictive ability, we are also interested in model interpretability, which variable contributes more to predicting systemic financial risks. In the XGBoost model, we can use gain, weight, and cover to represent feature relative importance. Among them, the gain is quite an important indicator that shows the average contribution of a certain feature to the prediction results in all decision trees. In other words, it represents the ability of a feature to split at a node, and the greater its ability to split, the more it contributes to ultimate results. The weight reflects the number of times a feature is used in all decision trees, and the more frequently used features contribute more to ultimate results. The cover indicates the impact of a feature on other features; if a feature can "cover" other features more, it has a more significant impact on ultimate results.

The gain, weight, and cover values for each variable in the model are shown in the table below, where weight is the average value derived by dividing the total number of decision trees:

Table 6 Gain, weight, and cover values for the XGBoost model

Variable	Gain	Weight	Cover
rgdp_growth	0.5435	0.5892	77.7736
Iy	0.5557	0.5066	80.1890
gl/gdp	1.0116	0.6422	83.4319
cagdp	1.0051	0.7655	94.7718
Money	1.1391	0.5392	90.9355
stir	2.2947	0.9715	121.0461
ltir	0.8423	0.7238	88.3569
hpnom	2.4966	1.3252	125.3229
unemp	1.3163	0.9042	91.1716
debtgdp	1.9842	0.7288	109.3101
revenue	1.2668	0.5719	82.9016
expenditure	1.0608	0.2712	76.1590
tloan	1.2036	0.5545	86.8642
tmort	0.7334	0.3884	71.7516
thh	4.4437	0.5240	170.9257
bdebt	1.9352	0.2202	105.3754
eq_tr	1.8642	0.8338	105.7663
housing_tr	0.6042	0.6565	83.7297
bond_tr	0.7365	0.9154	77.5267
lev	2.4904	0.9419	125.9894
ltd	5.6543	0.8145	179.2764
noncore	1.3816	0.6748	109.6545

As shown in the table, the feature with the greatest gain value is the bank deposit-to-loan ratio, followed by household sector loans, and the third is housing prices. The gain value of the bank capital adequacy ratio is close to that of housing prices. This result is similar to the logic behind the US subprime crisis: banks issued a large number of subprime loans to households with poor repayment ability, leading to an increase in bank deposit-to-loan ratio and household sector debts; when housing prices fell, a large number of loans could not be repaid, and the crisis ensued. Similar conclusions were drawn from weight and cover values.

A major problem in using gain, weight, and cover values to observe feature importance is that when the marginal contribution of a certain feature changes, the contribution of other features also changes, which does not satisfy "consistency." Therefore, this article further adopts a permutation test and *SHAP* values to measure feature contribution.

The basic idea of the permutation test is to randomly reorganize (shuffle) the values of each feature in the training sample and then observe the effect of such an operation on classifier performance. If classifier performance decreases, the feature has a certain degree of contribution to the classification results; if classifier performance does not change significantly, the feature's contribution to the classification results is small. The importance scores of each feature obtained from the permutation test are depicted in the figure below:

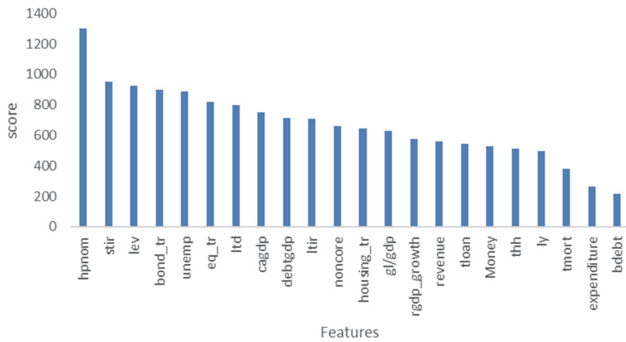

Figure 2 Feature importance based on the permutation method

As can be seen, the highest contribution still comes from housing prices, consistent with many early research findings. However, the differences between the importance scores of many other variables are relatively few. To refine the results, the *SHAP* (SHapley Additive Explanations) method is applied for further feature importance analysis.

The *SHAP* method is a feature importance evaluation method based on game theory, which can be used to interpret the prediction results of any tree-based model, including decision trees, random forests, and XGBoots. The *SHAP* method relates the feature importance to Shapley values in game theory and is an interpretable feature importance evaluation method. Moreover, the *SHAP* method guarantees local accuracy and has good consistency.

The *SHAP* method primarily uses *SHAP* values to estimate feature contribution, which means the average marginal contribution of feature values in all possible subsets of feature combinations, namely the weighted average of marginal contribution. *M* samples are extracted from each set *S*, and features other than set *S* are replaced by the extracted samples during each iteration to calculate the expected value of the model prediction results based on the current feature subset *S*. The *SHAP* value calculation formula for each feature value *i* is expressed as:

$$SHAP_i = \frac{1}{M}\sum_{m=1}^{M}\left(f\left(x_{+i}^m\right) - f\left(x_{-i}^m\right)\right) \tag{12}$$

where $f\left(x_{+i}^m\right)$ and $f\left(x_{-i}^m\right)$ are the prediction results of sample feature values, the feature values in $f\left(x_{+i}^m\right)$ are not replaced, but the feature values in $f\left(x_{-i}^m\right)$ are randomly replaced. The model *SHAP* importance is illustrated in Fig. 3:

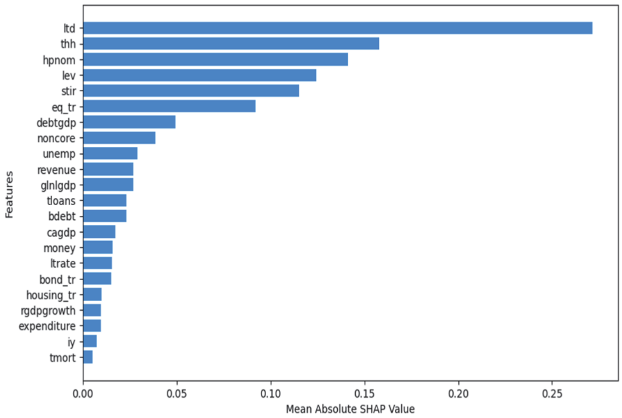

Figure 3 Feature importance based on the *SHAP* method

According to the feature importance results based on the *SHAP* method, the bank deposit-to-loan ratio contributes the most to predicting systemic financial risks, followed by household sector loans, housing prices, bank capital adequacy ratio, short-term interest rates, and stock returns. The feature importance of the remaining variables is relatively low. Particularly, both indicators, current account/GDP and annual growth rate of real GDP, mentioned by Tolo, have low feature contributions [29].

Meanwhile, we note that the previous literature usually used housing price growth rates as the predictive feature [31]. However, this article finds that the predictive contribution of gross property returns, including housing price growth rates, to systemic financial risks is much lower than the absolute level of housing prices, which may be because, as important collateral, the value of real estate itself is the most significant factor affecting financial risks; thus, its absolute level is more indicative of actual risks in the financial sector.

To distinguish how these features influence the prediction, a heat map of feature *SHAP* values is drawn (Fig. 4).

In the figure, darker colors represent larger *SHAP* values and greater feature values. We can see that a higher deposit-to-loan ratio positively impacts systemic financial risk prediction; in other words, it tends to increase the probability of systemic financial risk occurrence, and the other variables predict a similar direction for systemic risks. One difference is stock returns, as shown in the graph, which positively impact systemic financial risk prediction when they are at a lower level; that is, lower stock returns increase the probability of systemic financial risk occurrence.

One concern is that a higher bank capital adequacy ratio in the previous year has a positive impact on the occurrence of systemic financial risks, which may be explained by the fact that banks with higher capital adequacy ratios are more resilient to risks and face less regulatory pressure on capital adequacy. Hence, they are more likely to expand their lending business as the economic outlook changes. For instance, when housing prices increase, banks may issue more mortgages or real estate development loans and take on higher risks, thus resulting in an outbreak of systemic risks. It also tells us that raising the bank capital adequacy ratio is not a perfect way to prevent systemic financial risks.

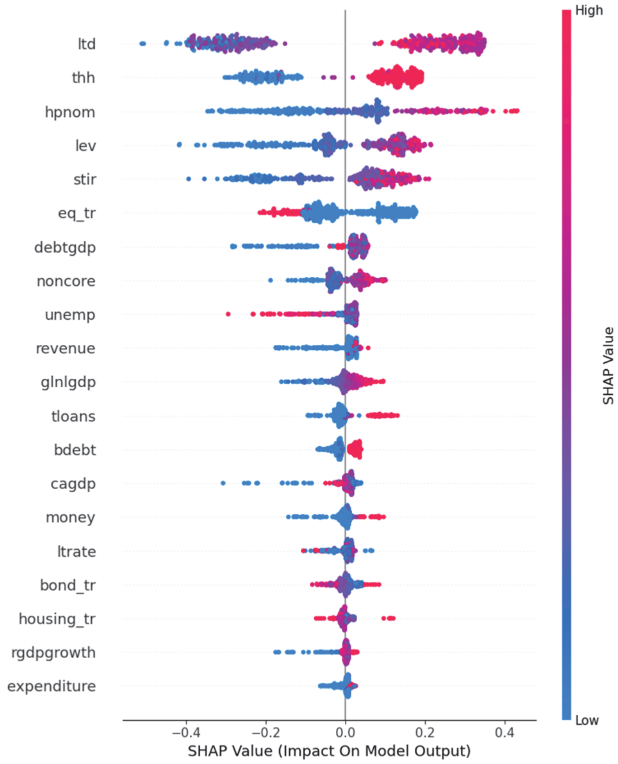

Figure 4 A heat map of *SHAP* values

4.5.3 Feature Dependency Analysis of Variables

Given the special role real estate has played in economic development and financial crises, as well as the information spillover effects of real estate on financial markets, this article further plots dependency graphs of *SHAP* values between housing prices and household sector debts, corporate sector debts, bank capital adequacy ratio, and bank deposit-to-loan ratio [35].

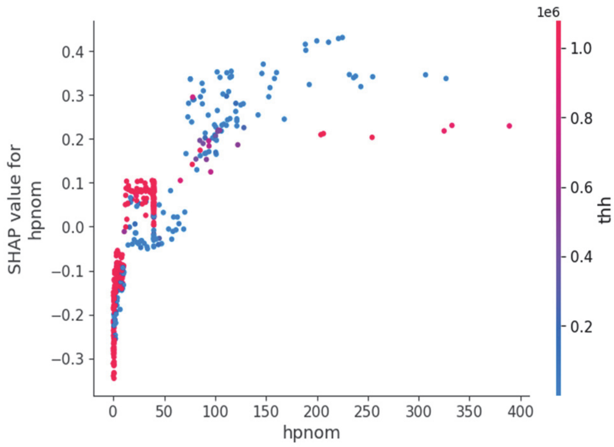

Figure 5 The *SHAP* dependency graph between housing prices and household debts

As shown in Fig. 5, higher housing prices are more likely to positively affect the debt levels of low-debt households (blue) and increase their debt levels, while the effect on high-debt households is less certain. As real estate is common collateral, it is easy to understand that higher collateral values will increase loan availability for low-debt households, similar to the early stages of the subprime crisis. Combined with the previous global *SHAP* value analysis, the importance of household debts in predicting systemic financial risks is second only to the bank deposit-to-loan ratio. It means that if the overall household debts in society are low, higher housing prices are prone to cause systemic financial risks.

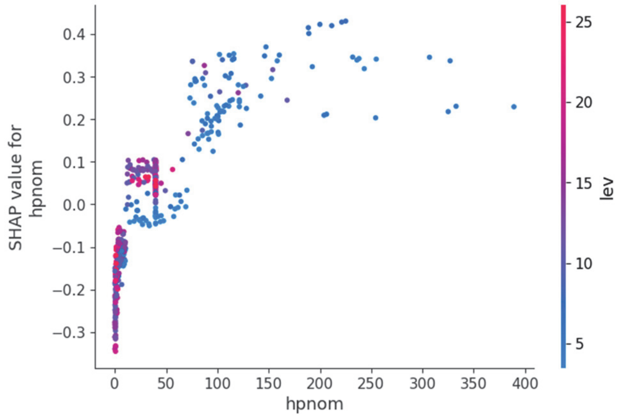

Figure 6 The *SHAP* dependency graph between housing prices and corporate debts

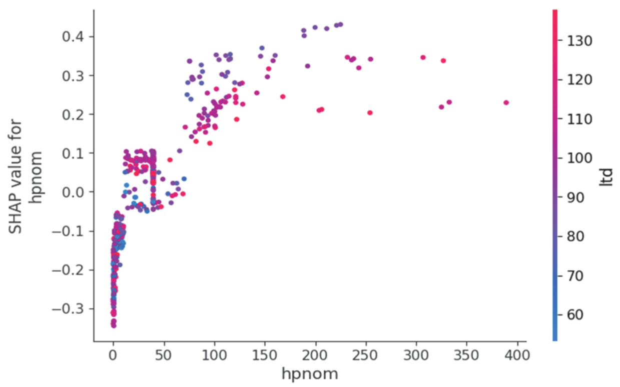

Figure 7 The *SHAP* dependency graph between housing prices and the bank capital adequacy ratio

Fig. 6 shows a similar relationship between housing prices and corporate debts. The above analysis demonstrates that "housing not for speculation" is still necessary. When the debts of both households and businesses are low, housing prices pushed up by a few speculators may lead to a rapid rise in the debts of the entire society, increasing the probability of systemic financial risk occurrence.

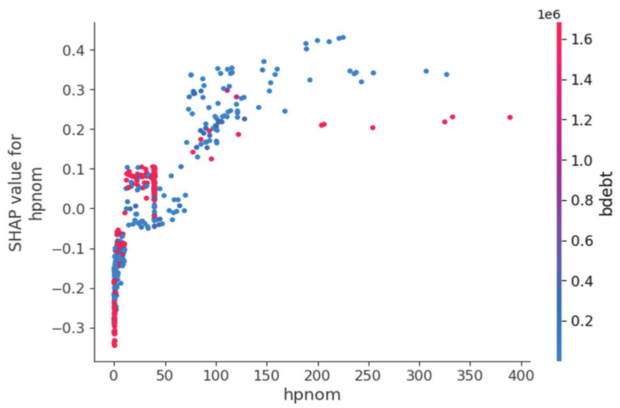

Figure 8 The *SHAP* dependency graph between housing prices and the bank deposit-to-loan ratio

Fig. 7 illustrates the relationship between housing prices and the bank capital adequacy ratio. For banks with a capital adequacy ratio at low to medium levels, housing prices increase their capital adequacy ratio; for banks with a capital adequacy ratio at high levels, housing prices negatively affect their capital adequacy ratio. It may be because banks with a higher capital adequacy ratio face

Less regulatory pressure on capital adequacy and are more resilient to risks, incentivizing them to issue more mortgages. When housing prices rise, more mortgages cause an increase in weighted risky assets, accruing risks. A chain reaction ensues once the risks break out, which echoes the previous global *SHAP* importance analysis. Therefore, the capital adequacy ratio should not be regarded statically, and banks with a high capital adequacy ratio should be rigorously and dynamically monitored during rising housing prices.

Fig. 8 indicates the *SHAP* dependency between housing prices and the bank deposit-to-loan ratio. Obviously, for banks with a deposit-to-loan ratio at medium to high levels, housing prices positively affect their deposit-to-loan ratio in a significant way. In combination with the previous ranking of global *SHAP* value importance, it signals a higher probability of systemic financial risk occurrence.

This paper further calculates different subjects' systematic financial risk score to increase the model's practicability. The score for each subject is determined by calculating the logarithm of its systemic financial risk occurrence odds, namely:

$$Score = \log_2\left(P/1-P\right) \tag{13}$$

where P is the odds of systemic financial risk occurrence for a certain subject predicted by the model, the larger this value is, the higher the likelihood of systemic financial risk occurrence, so the greater the subject's risk.

A benchmark point, the 5% quantile of the model test set, is set as a reference. Each subject's systemic financial risk score is the difference between its log odds value and the log odds value at the benchmark point, namely:

$$Score = \left(\log_2\left[odds_i\right] - base\right) \cdot 50 \tag{14}$$

The score has been expanded by a factor of 50, but this does not change its meaning. We can consider the score's reasonableness from two perspectives. The first is accuracy, the proportion of the samples where systemic financial risks occur among the samples above a certain quantile. The second is coverage, the proportion of samples above a certain quantile where systemic financial risks occur, to the total number of samples where systemic financial risks occur. For the data in this paper, the two indicators above the 5%, 10%, and 20% quartiles respectively are shown in the table:

Table 7 Quantile, *accuracy*, and coverage

Quantile	*Accuracy*	Coverage
5%	0.8696	0.2273
10%	0.9130	0.4773
20%	0.9121	0.9432

The table shows that the samples with systemic financial risk scores above 20% quantile can cover more than 94% of the samples with systemic financial risks. Its *accuracy* rate reaches 91.21%. Therefore, the systemic financial risk score constructed in this paper can be a reliable early warning indicator for systemic financial risks.

5 CONCLUSIONS AND RECOMMENDATIONS

Systemic financial risks are highly concealed and destructive, but an accurate and perfect measurement and early warning system has yet to be formed. Based on the XGBoost algorithm, this article proposes a framework for analyzing and predicting systemic financial risks and explores the application of machine learning to systemic risk prevention and control. This article combines historical data from the Jorda-Schularic-Taylor macroeconomic database with existing research to select 22 feature indicators, eliminates outliers using *LOF* and IF detection, and then constructs a prediction model with the XGBoost algorithm. The main conclusions and recommendations are as follows:

First, the systematic financial risk prediction model trained by XGBoost significantly outperforms the traditional linear logistic model, and its *accuracy* rate, *AUC*, *recall* rate, and *precision* rate reach 84.77%, 91.49%, 85.23%, and 85.41%, respectively. Therefore, in preventing systemic financial risks and constructing early warning systems, we should focus on more than just the linear relationship between systemic financial risks and economic variables but should further explore the application of non-linear models, such as machine learning, and simultaneously take advantage of higher-order data characteristics.

Second, this article conducts a model interpretability analysis based on the gain, weight, and cover values as well as the *SHAP* method. In general, the bank deposit-to-loan ratio, household sector loans, housing prices, the bank capital adequacy ratio, short-term interest rates, and stock returns have good predictive ability for systemic financial risks, among which, except for a few features such as stock returns, the remaining variables have a positive impact on the probability of systemic financial risk occurrence. Housing prices, household debts, the bank deposit-to-loan ratio, the capital adequacy ratio, and the generation path of systemic financial risks remain the primary logic. As a pillar industry of the national economy, when the economy continues to be weak after the impact of the New Crown Epidemic, the real estate industry needs to take on the role of stimulating the economy. However, we should not underestimate the possibility of systemic financial risk occurrence arising from this, and the overall principle is still to adhere to "housing not for speculation," as well as moderate debts among real estate enterprises to prevent systemic financial risks.

Third, contrary to previous research findings, this article finds that a higher capital adequacy ratio in the current period signals a higher probability of future systemic financial risks. It suggests that monitoring capital adequacy ratios is not a foolproof measure to prevent systemic financial risks; in other words, a high capital adequacy ratio cannot ensure systemic financial risks will not arise. Therefore, we should view the process dynamically and establish a more dynamic and diversified supervision method and indicator systems to avoid excessive inflow of bank funds into high-risk areas through various channels. With the exploratory application of blockchain and other technologies in banking, this problem will likely become more accessible to solve [36].

Fourth, the real estate *SHAP* dependency analysis indicates that when household sector debts are low, one should be wary of the potential risks brought about by rising housing prices. Therefore, the risk level should not be judged statically just because the macro leverage ratio is not high, but when the overall macro leverage ratio is not high, especially when the leverage ratio of the household sector is not high, the phenomenon of individual speculation in real estate should be strictly controlled to prevent the overall leverage ratio from rising due to the boom in the real estate market. It can be achieved by the introduction of policies such as a property tax to reduce the number of properties held by individuals and an appropriate increase in the supply of real estate so that housing prices can be maintained in a reasonable and moderate range without rising or falling too rapidly, thus keeping the overall stability of macro leverage ratio, and preventing the occurrence of systemic financial risks.

Acknowledgments

This research project is supported by the Science Foundation of Beijing Language and Culture University (supported by "the Fundamental Research Funds for the Central Universities") (approval number 23YJ090001).

6 REFERENCES

[1] Caggiano, G., Calice, P., Leonida, L., & Kapetanios, G. (2016). Comparing logit-based early warning systems: Does the duration of systemic banking crises matter? *Journal of Empirical Finance, 37*, 104-116. https://doi.org/10.1016/j.jempfin.2016.01.005
[2] Wang, Y. L., Zheng, X. Y., Yin, X. M., & Cai, J. R. (2022). Simulation of Financing Decisions with Behavioural Preferences and Yield Uncertainty. *International Journal of Simulation Modelling, 21*(4), 675-683. https://doi.org/10.2507/IJSIMM21-4-CO16
[3] Ristolainen, K. (2018). Predicting banking crises with artificial neural networks: The role of nonlinearity and heterogeneity. *The Scandinavian Journal of Economics, 120*(1), 31-62. https://doi.org/10.1111/sjoe.12216
[4] Bluwstein, K., Buckmann, M., Joseph, A., Kapadia, S., & Şimşek, Ö. (2023). Credit growth, the yield curve and financial crisis prediction: Evidence from a machine learning approach. *Journal of International Economics, 145*(6). https://doi.org/10.1016/j.jinteco.2023.103773
[5] Zhou, W., Zhai, X., & Tan, H. (2022). Research on Financial frauds prediction model of Chinese public companies with XGBoost. *Journal of Quantitative & Technological Economics, 39*(7), 176-196.
[6] Rudin, C. (2018). Please stop explaining black box models for high-stakes decisions. *Stat, 1050*, 26.
[7] Beutel, J., List, S., & von Schweinitz, G. (2019). Does machine learning help us predict banking crises? *Journal of Financial Stability, 45*(6). https://doi.org/10.1016/j.jfs.2019.100693
[8] Kaminsky, G. L. & Reinhart, C. M. (1999). The twin crises: the causes of banking and balance-of-payments problems. *Journal of Economic Literature, 89*(3), 473-500.

https://doi.org/10.1257/aer.89.3.473
[9] Christensen, I. & Li, F. (2014). Predicting financial stress events: A signal extraction approach. *Journal of Financial Stability, 14*(5), 54-65. https://doi.org/10.1016/j.jfs.2014.08.005
[10] Davis, E. P. & Karim, D. (2008). Comparing early warning systems for banking crises. *Journal of Financial Stability, 4*(2), 89-120. https://doi.org/10.1016/j.jfs.2007.12.004
[11] Barrell, R., Davis, E. P., Karim, D., & Liadze, I. (2010). Bank regulation, property prices and early warning systems for banking crises in OECD countries. *Journal of Banking & Finance, 34*(9), 2255-2264. https://doi.org/10.1016/j.jbankfin.2010.02.015
[12] Filippopoulou, C., Galariotis, E., & Spyrou, S. (2020). An early warning system for predicting systemic banking crises in the Eurozone: A logit regression approach. *Journal of Economic Behavior & Organization, 172*(4), 344-363. https://doi.org/10.1016/j.jebo.2019.12.023
[13] Šošić, M. & Graovac, J. (2022). Effective methods for email classification: Is it a business or personal email? *Computer Science and Information Systems, 19*(3), 1155-1175. https://doi.org/10.2298/CSIS220212034S
[14] Rogic, S. & Kascelan, L. (2021). Class balancing in customer segments classification using support vector machine rule extraction and ensemble learning. *Computer Science and Information Systems, 18*(3), 893-925. https://doi.org/10.2298/CSIS200530052R
[15] Ibrahim, J. & Gajin, S. (2022). Entropy-based network traffic anomaly classification method resilient to deception. *Computer Science and Information Systems, 19*(1), 87-116. https://doi.org/10.2298/CSIS201229045I
[16] Shi, C., Pei, Y., Li, D., & Wu, T. (2021). Influencing factors of catering O2O customer experience: An approach integrating big data analytics with grounded theory. *Tehnički vjesnik, 28*(3), 862-872. https://doi.org/10.17559/TV-20210124041130
[17] Kurniawan, Y., Putra, A. B. Y., & Cahyadewi, N. P. (2023). Get to know P2P lending and investors learning process at Indonesia. *Journal of System and Management Sciences, 13*(1), 241-265. https://doi.org/10.33168/JSMS.2023.0114
[18] Li, D., Na, S., Ding, T., & Liu, C. (2021). Credit risk management of P2P network lending. *Tehnički vjesnik*, 28(4), 1145-1151. https://doi.org/10.17559/TV-20200210110508
[19] Chen, T., Cao, Z., & Cao, Y. (2021). Comparison of job satisfaction prediction models for construction workers: CART vs. Neural network. *Tehnički vjesnik, 28*(4), 1174-1181. https://doi.org/10.17559/TV-20200507022306
[20] Manasse, P. & Roubini, N. (2009). 'rules of thumb' for sovereign debt crises. *Journal of International Economics, 78*(2), 192-205. https://doi.org/10.1016/j.jinteco.2008.12.002
[21] Savona, R. & Vezzoli, M. (2015). Fitting and forecasting sovereign defaults using multiple risk signals. *Oxford Bulletin of Economics and Statistics, 77*(1), 66-92. https://doi.org/10.1111/obes.12052
[22] Alessi, L. & Detken, C. (2018). Identifying excessive credit growth and leverage. *Journal of Financial Stability, 35*, 215-225. https://doi.org/10.1016/j.jfs.2017.06.005
[23] Joy, M., Rusnák, M., Smidkova, K., & Borek, V. (2017). Banking and currency crises: differential diagnostics for developed countries. *International Journal of Finance & Economics, 22*(1), 44-67. https://doi.org/10.1002/ijfe.1570
[24] Casabianca, E. J., Catalano, M., Forni, L., Giarda, E., & Passeri, S. (2019). An early warning system for banking crises: from regression-based analysis to machine learning techniques."*Marco Fanno" Working Papers, 235*.
[25] Fouliard, J., Howell, M., Rey, H., & Stavrakeva, V. (2020). Answering the queen: Machine learning and financial crises. *National Bureau of Economic Research, 28302*.
[26] Kim, K. J. (2003). Financial time series forecasting using support vector machines. *Neurocomputing, 55*(1/2), 307-319. https://doi.org/10.1016/S0925-2312(03)00372-2
[27] Lee, Y. C. (2007). Application of support vector machines to corporate credit rating prediction. *Expert Systems with Applications, 33*(1), 67-74. https://doi.org/10.1016/j.eswa.2006.04.018
[28] Ahn, J. J., Oh, K. J., Kim, T. Y., & Kim, D. H. (2011). Usefulness of support vector machine to develop an early warning system for financial crisis. *Expert Systems with Applications, 38*(4), 2966-2973. https://doi.org/10.1016/j.eswa.2010.08.085
[29] Tl, E. (2020). Predicting systemic financial crises with recurrent neural networks. *Journal of Financial Stability, 49*(4), 100-146. https://doi.org/10.1016/j.jfs.2020.100746
[30] Chen, T. & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Knowledge Discovery and Data Mining. ACM*, 785-794. https://doi.org/10.1145/2939672.2939785
[31] Tl, E., Laakkonen, H., & Kalatie, S. (2018). Evaluating indicators for use in setting the countercyclical capital buffer. *International Journal of Central Banking, 14*, 51-111.
[32] Yue, D. (2008). Characteristics and detection for accounting fraud pattern of listed companies in China. *Tianjin China: Tianjin University Press*.
[33] Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 93-104. https://doi.org/10.1145/335191.335388
[34] Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. *2008 Eighth IEEE International Conference on Data Mining*, 413-422. https://doi.org/10.1109/ICDM.2008.17
[35] Choi, C. S. (2020). A study on the information spillover effects between housing and financial markets. *Journal of Logistics, Informatics and Service Science, 7*(2), 14-24. https://doi.org/10.33168/JLISS.2020.0202
[36] Hashem, A. (2023). The impact of blockchain technology on financial services: mediating role of big data analytics. *Journal of Logistics, Informatics and Service Science, 10*(2), 91-107. https://doi.org/10.33168/JLISS.2023.0207

Contact information:

Yunsong XU, PhD, Associate Professor
School of Business, Beijing Language and Culture University,
No. 15 Xueyuan South Road, Haidian District, Beijing, China
E-mail: xuyunsong@blcu.edu.cn

Jiaqi LI, PhD
(Corresponding author)
School of Finance, Central University of Finance and Economics,
No. 39 Xueyuan South Road, Haidian District, Beijing, China
E-mail: ljq4937@126.com

Anqi WU, Master Student
School of Business, East China University of Political Science and Law,
No. 555 Longyuan Road, Songjiang District, Shanghai, China
E-mail: Wuanqi61@163.com