Systematic Risk Prediction in Commercial Banks Based on Random Forest and BP Neural Network

Junbin Zhang[1,a], Peiying Zhang[1,b], Shiyang Song[2, c*], Junyu Su[3,d], Jinhai Tang[2,e]

f20092100174@cityu.mo[a], zpy1701@163.com[b], *Corresponding author: 2463756962@qq.com[c], sjy2166@163.com[d], 1992649158@qq.com[e]

Faculty of Finance,city University of Macau, Macau, China[1]
Alibaba Cloud Big Data Application College, Zhuhai College of Science and Technology, Zhuhai, China[2]
Faculty of Data Science, City University of Macau, Macau, China[3]

Abstract: Since the 1990s, the frequent occurrence of systemic financial risks culminating in financial crises has had a serious impact on the economies and financial systems of all countries. Systemic risk analysis has become a very important task for most central banks in the wake of the global financial crisis (GFC). The sudden and destructive nature of systemic financial risks requires that we should pay attention to the foresight of systemic financial risks. In this study, based on establishing a system of systemic financial risk characteristics indicators in China, we construct machine learning models of random forest and support vector machine to warn systemic financial risks in China, and compare the warning effects of the two models using confusion matrix, ROC curve (Receiver Operating Characteristic Curve) and dynamic warning analysis, and The main factors that drive up the level of systemic financial risk in China are identified.

Keywords: Random Forest, BP Neural Network, Risk Forecast, Machine Learning.

1 INTRODUCTION

Because indirect financing dominates in China and the banking sector plays an integral role in the overall economic system, the banking sector is the key area of significant financial risk in China. It is inevitable that instability in the banking sector will result in significant losses for other economic institutions. The country's systemic financial risk prevention should be focused on preventing major risks to banks. A major role in early warning of systemic financial risks has traditionally been played by traditional statistical and measurement methods.

Nowadays, with the development of computer parallel computing power and data science, the frontier technology of big data has gradually penetrated into various fields and brought revolutionary changes to the modern financial industry. Big data analysis methods represented by machine learning and deep learning have a series of advantages such as timeliness, accuracy, and sample size, which make them well suited for data analysis and information processing in the financial field and further enrich the means and tools for systemic financial risk early warning [4].

BDEDM 2023, January 06-08, Changsha, People's Republic of China

DOI 10.4108/eai.6-1-2023.2330362

In recessions, neuro-fuzzy models can improve the forecasting efficiency of daily stock market data, even though machine learning algorithms are still more effective during crisis periods [5]. A 45-year sample of banking systems in developed economies was analyzed to compare the out-of-sample forecasting ability of various early warning models, and machine learning algorithms were found to outperform Logit methods. A variety of complex patterns can be recognized by machine learning algorithms. In developing countries, machine learning algorithms can be used to enhance the efficiency of financial markets. The ability of machine learning algorithms to identify complex structures and make accurate predictions is attracting increasing attention from scholars. The need for intelligent risk warning systems is also high among industries and governments. For risk early warning research in financial markets, the use of machine learning algorithms will be crucial to building a risk control system.

Compared with the existing literature on systemic financial risk early warning, the contribution of this paper includes two main aspects: first, it adopts the frontier concept of "model uncertainty", and through a more standardized research model and analysis process, early warning of systemic financial risk of commercial banks is conducted based on big data methods such as random forest and BP neural network, and a better prediction model is found using model evaluation methods. [1]. Which improves the accuracy and effectiveness of systemic financial risk early warning in the banking industry to a large extent.

2 CURRENT RESEARCH STATUS

Systemic financial risk early warning techniques based on linear models are widely used in academia to detect economic crises and systemic financial risks. Research conducted by Chinese scholars using linear models such as FR has led to the establishment of an early warning system for major financial risks in China. To construct a capital market tail risk early warning system, an analysis of FR, STV, KLR and other early warning models was conducted [10]. In recent years, nonlinear models have replaced linear models in systematic financial risk forecasting and early warning. This is because linear models are unable to capture nonlinear relationships between economic and monetary variables, and they have enhanced early warning performance.

There has been an increasing amount of current research on the use of machine learning and deep learning nonlinear models to warn of systemic financial risks in foreign countries, and its analytical frameworks and analysis methods have also developed over time. In China, however, there is still relatively little research on the use of machine learning and deep learning to detect systemic financial risks early on. The purpose of this paper is to address the shortcomings of the existing literature by focusing on commercial banks and adopting machine learning and deep learning models for early warning analysis of systemic financial risks in China. The aim is to make early warning prediction of systemic financial risks more accurate and effective.

2.1 Random Forest

Random Forest is one of the more widely used and powerful machine learning methods, which is good at dealing with various types of prediction problems. Random forest is actually an integration of decision trees, which are usually trained by bagging method. Among them, a

decision tree is an algorithm that uses a tree structure to make decisions, consisting of a root node, several leaf nodes and internal nodes, where a leaf node represents a decision outcome and each other node represents an attribute test. A random forest is composed of several decision trees, and the best decision outcome is determined by voting on the decision outcome of each decision tree [2]. The basic idea of the random forest algorithm: k samples are drawn in the original training set using the bootstrap method with the same sample size as the original training set; k decision trees are built based on the k samples, resulting in k classification results; and each record in the k classification results is voted to determine its final classification.

2.2 BP Neural Network

BP neural network is a large nonlinear network, which simulates the human physiological reflex process. BP neural network simulates a large number of "neuron" nodes, in the input of a large amount of data, by itself to find the laws and logic between the nodes, and save the learning path between the nodes [7]. The concept of BP neural network was first introduced in the late 1980s, and its model building process is to explore the intrinsic connections and laws through the process of data input and output independently, without setting up the function relationship in advance. Through the gradient descent method, the BP neural network continuously adjusts its own weights, and when it does not reach a certain error accuracy, it reverses the stimulation until the final result meets the error accuracy.

The main idea of BP neural network is to estimate the error of the previous layer of the output layer based on the error after the output, and then use the error of this layer to estimate the error. The error of this layer is then used to estimate the error, so that the error estimates of all layers are obtained. The error estimate here can be understood as some kind of partial derivative, and we adjust the connection weights of each layer according to this partial derivative, and then recalculate the output error with the adjusted connection weights. BP neural network is a multilayer backward-looking intelligent network trained with error back propagation algorithm.

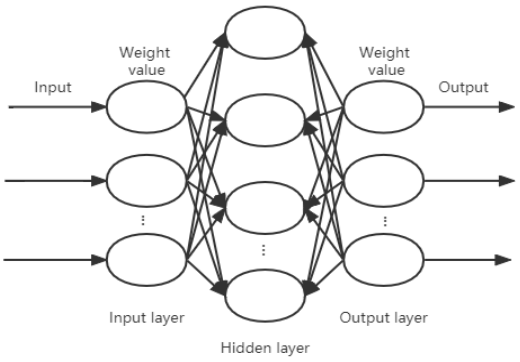


Figure 1 BP neural network model diagram

2.3 Commercial Banking Systemic Risk

The Financial Stability Board defines systemic risk as the risk of experiencing a strong

systemic event, which can occur only as a result of a "systemic event". It is caused by the loss of a financial institution and the exposure of an institution with "systemic importance" characteristics, the impact of which on other institutions cannot be underestimated. Systemic risk arises from the activities of banks, and when the relationship between banks gradually forms a network, one bank is affected and the whole banking system is widely affected [8]. The problem of systemic financial risk is not limited to the economic and financial spheres of a country, even if a country with a modest economy, once financial turmoil is generated it often affects the economic and financial situation in other parts of the world. In the accumulation phase, systemic financial risks have one of the most important characteristics - they accumulate in the upward financial cycle, are insidious and not easily identified. The existing study [6] used machine learning techniques such as K-nearest neighbors, random forests, support vector machines, Boosting, and regulatory assessment data on bank risks to build a UK bank crisis early warning system and compared it with traditional statistical techniques such as logistic regression, which found that machine learning techniques significantly outperformed traditional statistical techniques such as logistic regression, and in particular, random forests performed well and were suitable as a bank algorithmic model of the crisis early warning system.

3 RESEARCH METHODOLOGY

3.1 Feature Selection and Index System Construction

In terms of feature selection, in order to control the influence of individual bank characteristics and macroeconomic variables on the systemic risk of banks, based on the existing literature, the paper selects 11 control variables to establish a systemic financial risk indicator system for China's banking industry, taking into account the domestic and international research results and the actual national conditions of China.

Table 1 Commercial Bank Risk Measurement Variables Table

Variable symbol	Variable name	Variable definition
NII	Percentage of non-interest income	Non-interest income/total operating income
NIM	Net interest margin	Net interest income/Average balance of interest-bearing assets
NPL	non-performing loan ratio	Non-performing loans/total loan balances
ROE	return on equity	Net profit/average balance of assets
M2B	equity price-to-book ratio	Market value of the stock/book value
CRAR	capital adequacy ratio	Net capital/risk-weighted assets
SIZE	size of bank	Natural log of total bank assets
LEVERAGE	leverage ratio	Total liabilities/total assets
FD	Financial development	(stock market value + total bank credit)/GDP
M2_GR	Money supply growth	M2 quarterly year-on-year growth rate
GDP_GR	GDP growth	quarter-on-quarter GDP growth

3.2 Model Performance Evaluation Methods

3.2.1 Confusion matrix and ROC curve

Confusion matrix and ROC curve to evaluate the goodness of a model requires performance metrics, i.e., designing evaluation criteria to measure the generalization ability of the model. In machine learning and deep learning classification tasks, confusion matrix and ROC curve are more commonly used performance measures [9]. Among them, the confusion matrix is a more comprehensive representation of the model evaluation results, which classifies the predicted samples based on whether the true values are the same as the predicted values. One row of the confusion matrix is used to represent the true category of the sample, and one column is used to represent the predicted category of the sample. When the predicted category of the sample is the same as the true category of the sample, it means that the model predicts the correct classification of the sample. When the sample prediction category is different from the sample true category, it means the model predicts the wrong classification of the sample. The ROC curve is a common tool for analyzing the classification behavior of models with different thresholds.

Table 2 Confusion matrix example

True situation	Forecast result	
	Risk	Normal
Risk	TP (Real risk)	FN (false normal)
Normal	FP(False risk)	TN (True normal)

3.2.2 Cross-validation

In recent years, cross-validation has been used to select model weights in various model settings, including heteroskedasticity linear regression models, linear regression models with lagged dependent variables. Cross-validation is a statistical method commonly used in machine learning and deep learning to evaluate the generalization performance of models through experimental tests, which is more scientific and reasonable than the usual method of dividing data into training and testing sets in a single pass [6]. In the cross-validation process, the data is generally partitioned several times and multiple models are trained at the same time. A common cross-validation method is k-fold cross-validation, where k is an arbitrary number that can be specified.

First, the data set is randomly cut into k disjoint subsets of the same size; then k-1 subsets are used as training sets to train the model, and the remaining (held out) one subset is used as a test set to test the model; the previous step is repeated for the possible k choices (each time a different subset is picked as the test set); thus k models are trained, and the test error is calculated for each model on the corresponding test set to obtain k test errors, and a cross-validation error is obtained by averaging these k test errors

Figure 2 k-fold cross-validation (k-fold cross-validation) validation schematic

4 EXPERIMENTAL RESULTS

In this paper, we use data from 34 commercial banks in China, take 11 indicator variables in the systemic financial risk characteristic indicator system of China from January 2019 to December 2021 as the input of the early warning mode [3]. Take the sequence of risk early warning dummy variables obtained by transforming the results of risk monitoring analysis as the model expectation output, divide the data in a random single pass (75% of the data as the training set and 25% of the The data are divided randomly in a single pass (75% of the data as training set and 25% as test set), and the systematic financial risk early warning models of random forest and BP neural network are constructed in JupyterNotebook, which supports Python language, respectively. In this paper, 11 indicator variables in the systemic financial risk characteristic indicator system of China from January 2008-December 2017 are used as the input of the early warning model, and the sequence of risk warning dummy variables obtained by transforming the results of risk monitoring analysis is used as the expected output of the model, and the data are divided randomly in a single pass (75% of the data as the training set and 25% of the data as the test set), and the systemic financial risk early warning models of random forest and BP neural network are constructed in JupyterNotebook, which supports Python language, respectively.

Results of static warning accuracy and recall for random forest model, BP neural network model.

Table 3 Model alert accuracy and recall results

Model	Recall rate	Accuracy rate
Random forest	0.89	0.91
BP neural network	0.95	0.97

Through the results in Tables 3, it can be concluded that the early warning effect of the deep learning model of BP neural network is better than that of the random forest model, and the accuracy and recall rates of the two types of models on the training set are 0.91 and 0.97, 0.95

and 0.89, respectively. The bp neural network is better than the random forest model with an AUC value of 0.965932. The BP neural network model, whose ROC curve is closer to the upper left corner of the coordinate system than the model, has an AUC value of 0.986235.

Table 4:ROC curves of BP neural network

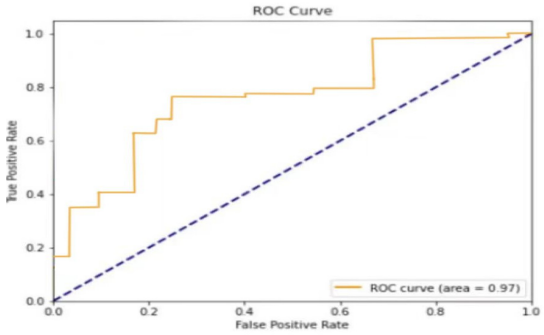
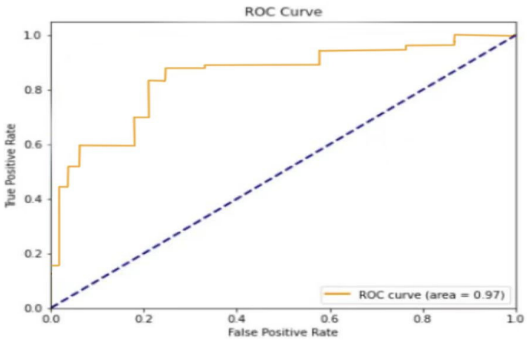


Table 5:ROC curves for random forests



5 CONCLUSION

This paper employs random forest and BP neural network deep learning models for systemic financial risk early warning in China, comparing and analyzing the early warning effects of different models, as well as comparing the differences in outcome estimation and prediction between traditional econometric models and machine learning and deep learning algorithm models, and identifying the main causes that push up the level of systemic financial risk in China.

The study concludes that, first, in terms of various performance measures and prediction results, machine learning and deep learning algorithmThe results of ROC curve model evaluation and cross-validation show that machine learning and deep learning algorithms significantly outperform traditional econometric models.The results of ROC curve model evaluation and cross-validation show that machine learning and deep learning algorithm

models perform better than traditional econometric models. In the future, in terms of systemic financial risk early warning, we will continue to optimize and improve the early warning of systemic financial risk in China

In the future, in terms of systemic financial risk early warning, we will continue to optimize and improve the early warning model of China's systemic financial risk, and introduce more cutting-edge big data analysis technologies such as transfer learning, Meta Learning, Explainable AI and other machine learning and artificial intelligence. to continuously improve the efficiency and performance of systematic In addition, we will revise the data and methods according to the actual needs, and build early warning models that can predict systemic financial risks over a longer period of time on the basis of guaranteeing the accuracy and credibility of prediction.

REFERENCES

[1] Beutel, J., List, S., & Schweinitz, G. (2019). Does machine learning help us predict banking crises. Journal of Financial Stability, 45, 1–28.

[2] Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25(2), 197-227.

[3] Dai Shugeng, Lin Binzhao & Yu Bo.(2022). Financial opening, bank connective degree and the systemic risk of our country bank. Jilin university journal of social sciences (05), 101-117 + 237. Doi: 10.15939 / j.j ujsse. 2022.05 jj2.

[4] Goodell, J. W., Kumar, S., Lim, W. M., & Pattnaik, D. (2021). Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. Journal of Behavioral and Experimental Finance, 32, 100577.

[5] Guresen, E., Kayakutlu, G., & Daim, T. U. (2011). Using artificial neural network models in stock market index prediction. Expert Systems with Applications, 38(8), 10389-10397.

[6] Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. Encyclopedia of database systems, 5, 532-538.

[7] JIN, Wen, et al. The improvements of BP neural network learning algorithm. In: WCC 2000-ICSP 2000. 2000 5th international conference on signal processing proceedings. 16th world computer congress 2000. IEEE, 2000. p. 1647-1649.

[8] Kaufman, G. G., & Scott, K. E. (2003). What is systemic risk, and do bank regulators retard or contribute to it?. The independent review, 7(3), 371-391.

[9] Yang, S., & Berdine, G. (2017). The receiver operating characteristic (ROC) curve. The Southwest Respiratory and Critical Care Chronicles, 5(19), 34-36.

[10] Zhang, Z., & Chen, Y. (2022). Tail risk early warning system for capital markets based on machine learning algorithms. Computational Economics, 60(3), 901-923.