Risk Prediction Method of Village Banks Based on Equity Relationship

Ailun Zhang(✉)

Jinan University, Guangzhou 510632, China
zhangailun368@gmail.com

Abstract. In the context of the simultaneous occurrence of financial risks in four rural banks in Henan, China, the public is worried about rural banks. Unlike big banks, which are "too big to fail", rural banks face the problem of "too many to fail". In this paper, an indicator system is established by selecting all the beneficial shareholders and their corresponding share data in the corporate equity of 82 banks. Based on the decision tree, KNN algorithm and RUSBoost algorithm, a risk early warning model for Henan rural banks is constructed, and it is verified that the three algorithms can be used to a certain extent. Two states of risk and normal were classified, and the RUSBoost model performed the best, with a false positive rate (FPR) value of 7% and a false positive rate (FPR) value of 0. The control group was introduced, and the equity data that was not processed by the index system was studied based on the same algorithm, and it was concluded that the index system had a good effect. This is the first attempt to identify the risk status of rural banks in my country by building an equity structure index system based on machine learning technology and will provide a prototype technical means and thinking direction for rural bank operators and regulatory authorities to provide timely early warning.

Keywords: illage bank · Machine learning · Risk early warning · Shareholding structure

1 Introduction

In April 2022, four rural banks in Henan Province closed online withdrawal and transfer channels at almost the same time, causing panic among depositors. The China Banking and Insurance Regulatory Commission said the case stemmed from collusion among the bank's shareholders, both inside and outside the bank. According to reports, at least 10 billion yuan of funds were involved, and about 400,000 depositors were affected.

By the end of 2021, there were 1,651 rural banks in mainland China, accounting for about 36% of the total number of banking financial institutions. Among them, there are the largest number of village and town banks in Shandong, Hebei and Henan, with 126, 110 and 86 respectively. According to statistics from the People's Bank of China, as of the second quarter of 2021, 122 village and township banks were classified as high-risk institutions, accounting for about 29% of all high-risk institutions. According

Y. Jiang et al. (Eds.): ICFIED 2023, AEBMR 237, pp. 339–349, 2023.
https://doi.org/10.2991/978-94-6463-142-5_38

to a research report released by Huaan Securities, the non-performing loan ratio of rural banks in mainland China will be as high as 4% in 2020, much higher than other financial institutions. my country has always attached great importance to the identification and prevention of financial risks. China's 19th National Congress put the prevention and resolution of major crises at the forefront of the "three tough battles". In the "14th Five-Year Plan", the State Council also clearly proposed to "protect financial security and keep the bottom line against the outbreak of systemic risks" [1]. Since China's current main investment method is still indirect investment, and banking institutions also play an important role in indirect investment, China's systemic financial risks are mainly concentrated in the commercial banking system. At the same time, under the realistic background that China's economy is facing the impact of the new crown pneumonia epidemic and the possibility and expectations of the world economic recovery are still uncertain, the pressure on the regulatory system of China's commercial banks is also increasing, and the operation of reducing market risks is also facing certain challenges. As one of the important financial institutions in China's rural areas, village banks have played an indispensable role in the economic development of rural areas. Compared with the "too big to fail" of big banks, village banks are faced with the problem of "too many to fail" [2]. Due to the large number of village banks, a good risk warning for village banks can effectively prevent the spread of systemic risks in the banking industry and building a risk warning system that conforms to the operating characteristics and regulatory framework of China's village banks will help strengthen the risk management level of small banks and preventing major financial risks is of great positive significance. Therefore, this paper uses several mainstream machine learning techniques, such as decision tree, K-nearest neighbor algorithm KNN, ensemble learning, etc., to try to give a certain degree of financial risk warning to village banks.

2 Literature Review

Since village banks are built in rural areas of China, they are banking institutions that mainly provide financial services for local farmers, agricultural production and rural social and economic development. It is essentially a rural community bank, so it can be studied with reference to the risk source of rural community banks. On the whole, the sharp increase in the operating risks of rural commercial banks in my country in recent years is mainly due to the following problems: First, the enterprise management is not sound, and there is no good way and basic principles for enterprise management (Zhou Xiaochuan, 2020) [3]; The risks mainly come from the internal control management risks caused by the diversification of the ownership structure, the potential intervention risks of shareholders and the operation risks of shareholders' funds occupying. Second, the main services and assets are relatively simple and highly dependent on interest income. Since 2019, especially during the COVID-19 epidemic, the monetary policy has been structurally loosened, and the profitability has slowed down. The source of income of small commercial banks has a significant impact. The third is to locate village banks in the financial service supply of special regions, focusing on serving local farmers, agricultural production and rural economic development. The time limit is very limited. But once the village bank is in crisis, it may lead to the occurrence of systemic financial

risk due to the huge quantity and risk mechanism [4]. Among the above reasons, the second and third reasons have general effects and are not within the scope of this paper. The first reason has specific effects, so this paper attempts to study from the data of the ownership structure and tries to explore whether these village banks with financial risks have identifiable commonalities in the ownership structure by means of machine learning.

Among the problems of ownership structure, the most important is the problem of ownership concentration [5]. In the problem of ownership structure research, the selection of quantitative indicators of ownership concentration is the key to analyzing the problem. There are three common quantitative indicators, namely Herfindahl index, CR index and Z index. The Herfindahl index is a measure of the company's shareholding structure, and its value is the sum of the squares of the shares held by the top shareholders in the company to express the company's share concentration. The CR index is the sum of the shares held by shareholders in the company. The Z-index is the ratio of the shares held by the major shareholder with the largest share and the major shareholder with the second largest share in a company. The construction of the equity index system in this paper will be based on these three indexes.

Foreign experts' research on enterprise risk early warning mode started earlier, the data used is rich, and the theoretical context is relatively complete and mature. The main theoretical method used by Altman (1968) [6] is the multi-discriminant analysis technique, which uses a multivariate Z-score model to achieve the purpose of identifying corporate bankruptcy. Lane et al. (1986) [7] estimated the time of bank failure through the Cox model. Chiaramonte et al. (2015) [8] compared the effects of two different models in judging bank financial risk. Ferriani (2019) [9] used the Logit model to determine the quantitative default probability of general banks, creating an early warning model with a forecast accuracy of 4–6 quarters. Suss and Treitel (2019) [10] invoked confidential regulatory data and risk regulatory assessment scores, and compared models such as random forest, Boosting technology, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and their combined models. Estimated level of loss.

Domestic experts have made some explorations in the field of early warning of business risks of commercial banks, but most of them adopt subjective and qualitative methods. There is no corresponding exploration in the field of early warning of business risks of small commercial banks, especially rural commercial banks, and the exploration of the application of machine learning in the field of business risk assessment is not perfect. In terms of experimental data, due to insufficient disclosure of monitoring data such as operational information by the banking industry, especially small and medium-sized banks, very little information is obtained from public channels. Therefore, we try to establish a risk early warning model for domestic rural commercial banks through a model based on real data structure, and introduce decision tree, KNN algorithm and RUSBoost algorithm to fill the gap in research methods.

3 Data Processing

The data in this article mainly comes from Qichacha, a small and medium-sized enterprise credit information agency registered in the official record, which mainly provides

data search for small and medium-sized enterprises in China, including business information search of small and medium-sized enterprises, personal credit information search, operation information and other related information. The research sample of this paper covers 82 rural banks in Henan, and the data content is all the beneficial shareholders and their corresponding shares in the enterprise equity of the 82 banks, and they are sorted according to their shareholding ratios from high to low.

The data will be divided into processed and unprocessed data sets, and the two sets of data will be compared for algorithm learning. It is used to study whether the ownership structure index system constructed in this paper can play a certain role in risk early warning.

The unprocessed data set consists of four features, which are the shareholding ratios of the top four shareholders of the bank.

The preprocessed data set will use the constructed equity structure indicator system. The discussion of the ownership structure is mostly carried out from the perspective of the concentration of the enterprise's equity and the shareholding ratio of the insiders, and we mostly take the concentration of the equity and the shareholding ratio of the insiders as the main measure of the ownership structure that restricts the performance of the enterprise. Therefore, we can choose three indexes, Herfindahl index, CR index and Z index, which reflect the company's share concentration and insider's shareholding ratio respectively. The Herfindahl index is a measure of the company's shareholding structure, and its value is the sum of the squares of the shares held by the top shareholders in the company to express the company's share concentration. Due to the limited number of shareholders of village banks, this paper selects the first four shareholders. The CR index is the sum of the shares held by shareholders in the company. It is a quantitative technical indicator that measures the aggregation or dispersion of shares. It is used to illustrate the degree of share aggregation. The data of the top four shareholders are also used here. As an evaluation index of equity concentration, the Z index is mainly used to reflect the degree of influence of the shareholders with the first share on the business operation or stock market behavior. The Z-index is the ratio of the shares held by the major shareholder with the largest share and the major shareholder with the second largest share in a company. The higher the Z-index, the greater the difference in weight between the major shareholder with the first share and the major shareholder with the second share. These three indexes are used as a feature of the model, and we introduce the fourth feature, whether the bank has state-owned capital participation, using binary classification, participation is 1, no participation is 0.

Table 1. Comparison feature matrix (Self-drawn)

Name	Largest shareholder	Second largest	Third largest	Fourth largest
sample	0.5819	0.1	0.0078	0.0078

Table 2. Model feature matrix (Self-drawn)

Name	Herfindahl	CR	Z	State-owned assets
sample	0.348463	0.697324	5.8167	1

Table 3. Misjudgment cost matrix (Self-drawn)

Predict Actual	Normal	Risky
Normal	0	1
Risky	2	0

4 Model

The machine learning algorithm in this paper is set as follows. First, binary classification is used. If the bank cannot carry out basic business such as cash withdrawal due to cash flow breakage and other reasons, it is considered that the bank is in a risky state, and it is marked as 1, otherwise it is 0. Second, considering that the relevant government agencies usually adopt a more conservative mentality, they would rather issue false warnings than omit the risk situation of banks and further aggravate the risk situation. Therefore, this paper determines the asymmetric cost of wrong views, and puts the cost of type 1 error (omission of risky banks) is set to be twice the cost of type 2 errors (misreported risky banks) (Brauning et al., 2019) [11].

Divide all data types into training set and test set according to the ratio of 3:1, use the training set to train the model, and then use the test set data to test, the AUC value and confusion matrix of the training model to the test set can be obtained.

4.1 Decision Tree

Decision tree is one of the most widely used analytical model methods in the field of computer learning. The development process of the decision tree is essentially to classify the different traits of the training samples into different types of development processes (that is, branches) one by one, and select the one with the best result from various possible branches. The quality of the bifurcation and the degree of goodness or badness of the predicted results are determined by the purity of the random variables in the node, which is measured by the message entropy of the random variables in the node. The higher the uncertainty, the higher the information entropy, and the lower the purity.

Through a series of feature selection and decision rules, the node branches to the last classification leaf node, and the final output conclusion of classification calculation through recursion is the next decision tree. For our expectation and empirical analysis method, the process starts from the root node, and creates a route according to the values

of the series of prediction indicators until it reaches the leaf node, and finally determines whether the banking system has reached the risk status.

4.2 K-nearest Neighbor Algorithm

The KNN method is the K-nearest neighbor algorithm, which was first proposed by Cover and Hart in 1968, a method that has been quite perfect in theory. The basic idea of this analysis method is to assume that most of the K closest data in the feature space (expressed as being the closest in the feature space) are in a certain type, and the data is in these kinds of. In terms of type selection, the analysis method only judges the type it is in by comparing the types of the most adjacent one or several data types.

The KNN algorithm is especially suitable for the automatic analysis of data sets with very large data volume but using this method in some data sets with very small data

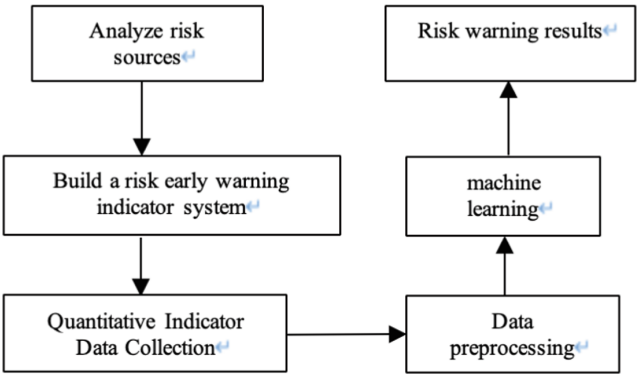

Fig. 1. Construction process of risk early warning model (Self-drawn)

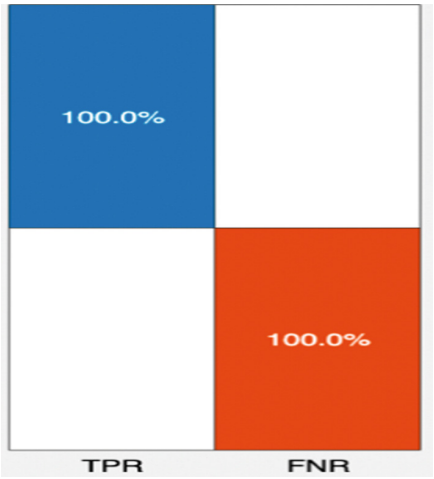

Fig. 2. Confusion matrix TPR and FNR of the control group (From MATLAB)

volume is particularly prone to bias. Due to the small sample size of village banks, the expected performance is not stable when dealing with village bank data.

4.3 RUSBoost

The above two algorithms are widely applicable, but for data imbalance classification, that is, the classification of data sets with large differences in the number of samples between categories is prone to errors. The specific performance in the data is that the samples of risky banks are very different from normal banks. Therefore, the RUSBoost algorithm is introduced.

RUSBoost is a very simple method for uneven datasets, which can be decomposed into RUS and Boost. RUS (random under sampling) refers to under sampling, that is, randomly selecting many multi-class data from the data set and forming a training sample set with a balanced layout with individual classes. Boost refers to the Adaboost.M2 algorithm. RUSBoost is to extract the training results through the RUS algorithm before using the weak classifier in each iteration of the Adaboost.M algorithm, and then perform the weak classifier exercise.

5 Model Results

Due to the extremely unbalanced distribution of the two types of samples, the results of focusing on the prediction accuracy are misleading, so the data of the false positive rate and false negative rate inside and outside the sample are introduced. The false negative rate (FPR) is the ratio of the number of negative sample outcomes predicted to be positive to the actual number of negative samples. The value of the false positive rate (FPR) is the ratio of the number of positive sample outcomes predicted to be negative to the actual number of positive samples. The two can well reflect the incidence of Type I errors (omission of risky banks) and Type II errors (misreported risky banks). From a regulatory point of view, regulators would rather misreport risks than underreport risks, and the former has a higher priority than the latter Figs. 1, 2, 3 and 4.

5.1 Control Group Training Results

Observing the data, it can be found that the false negative rate (FPR) of the decision tree, KNN and RUSBoost algorithms in the test data is 100%, which means that the data

Table 4. Model training results (Self-drawn)

	Train set accuracy	Test set accuracy	Train set false negative rate	Train set false positive rate	Test set false negative rate	Test set false positive rate
Decision tree	93.30%	95.50%	33.30%	5.30%	0%	4.80%
KNN	98.30%	100%	33.30%	0%	0%	0%
RUSBoost	93.30%	86.40%	0%	7%	0%	14.30%

that has not been processed by the model data has almost no ability to identify risks in machine learning.

5.2 Model Training Results

The decision tree uses Bayesian optimization to obtain an optimal number of splits of 11 through 30 iterations of expected improvement per second. The splitting criterion is to minimize the deviation. The prediction accuracy of training set is 93.3%, and the prediction accuracy of test set is 93.3%. The training set false-negative rate was 33.3%,

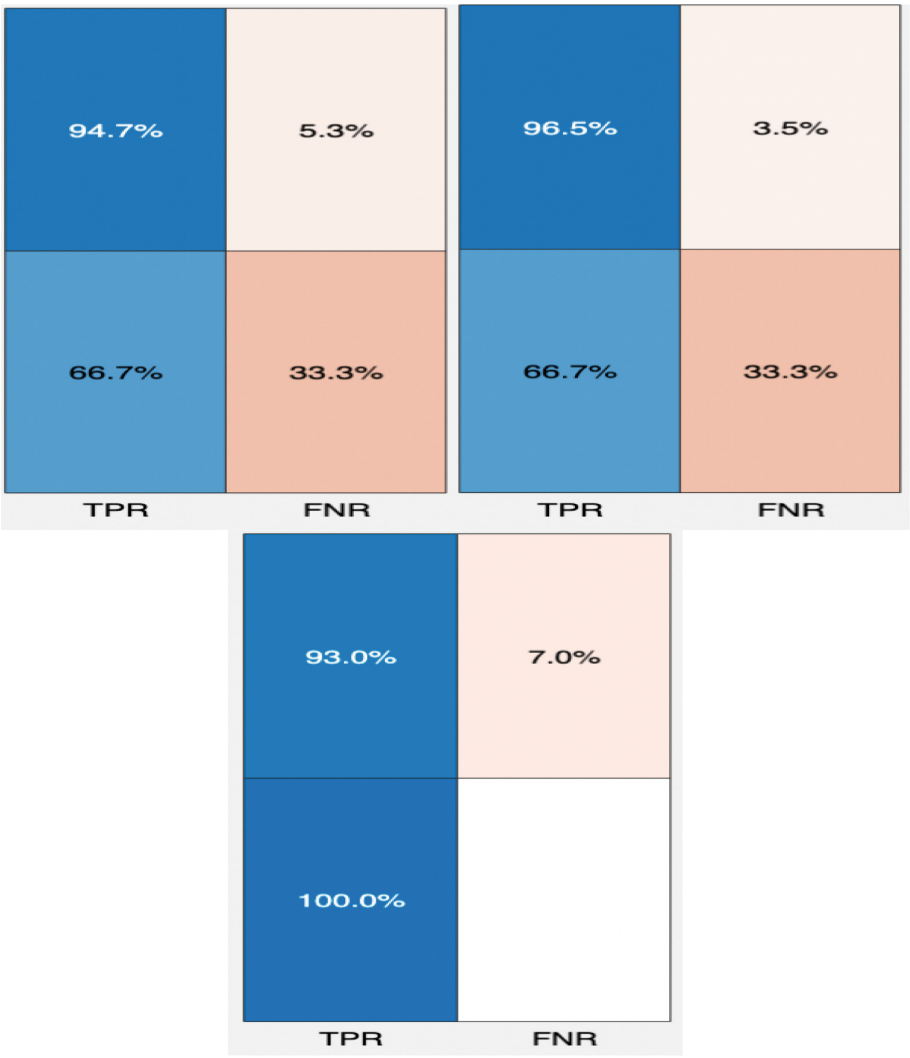

Fig. 3. Confusion matrices TPR and FNR of the training group (From MATLAB)

the training set false-positive rate was 5.3%, the test set false-negative rate was 0%, and the test set false-positive rate was 4.8%.

KNN uses Bayesian optimization to determine the hyperparameters through 30 iterations of expected improvement per second, the optimal number of neighbors is 12, the distance metric adopts Chebyshev, and the distance weight adopts the inverse distance weight, and the prediction accuracy of the training set is 98.3%. The prediction accuracy of test set is 100%, the training set false-negative rate is 33.3%, the training set false-positive rate is 0%, the test set false-negative rate is 0%, and the test set false positive rate is 0%.

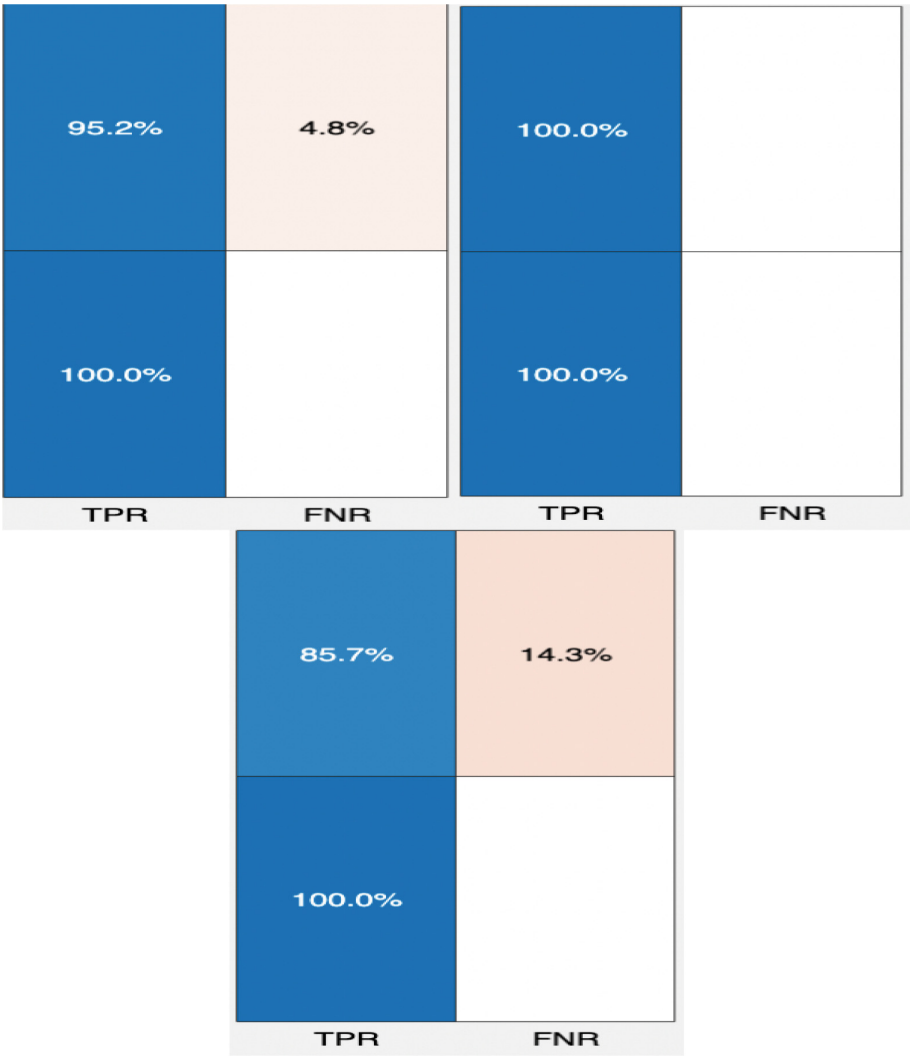

Fig. 4. Test group confusion matrix TPR and FNR (From MATLAB)

In the integrated learning tool RUSBoost, the maximum number of splits after Bayesian optimization is 1, the number of learners is 320, and the learning rate is 0.01017, the prediction accuracy of training set data is 93.3%, and the prediction accuracy of test set is 86.4%, the training set false-negative rate is 0%, the training set false-positive rate is 7%, the test set false-negative rate is 0%, and the test set false-positive rate is 14.3%.

As far as the prediction results are concerned, the RUSBoost algorithm has good accuracy in both training set and test set in risk detection. The KNN algorithm has poor accuracy in training set data prediction but has good performance in test set detection. The decision tree method has relatively poor ability to predict risk Tables 1, 2, 3 and 4.

6 Conclusion

The data source of this paper is the data of all beneficial shareholders and their corresponding shares in the corporate equity of 82 rural banks in Henan. This paper uses decision tree, KNN algorithm and RUSBoost algorithm to build a risk early warning model for rural banks in Henan. It is verified that the three algorithms can be classified to a certain extent. The two states of risk and normal provide a rudimentary technical means and thinking direction for village bank operators and supervisory departments to provide timely early warning. Among the three algorithms, the RUSBoost algorithm has the highest stability and better performance.

The limitation of this paper is that the amount of sample data is small, and a more complete model cannot be obtained. The model only uses the data in Henan Province, and its applicability to other provincial data is unknown. At the same time, there is a large difference between the minority class samples and the majority class samples in the sample, and the sampling method has a great influence on the prediction results. This problem also needs to be improved in future work.

References

1. Zhan Xiangyang, Zheng Yanwen. Keeping the Bottom Line of No Systemic Risks [J]. Chinese Financiers,2016 (10): 127-128.
2. Xu Chao. “Too Big to Fail” Theory: Origin, Development and Controversy [J]. International Finance Research,2013(8): 89-96.
3. Zhou Xiaochuan. Corporate Governance and Financial Stability [J]. China Finance,2020 (15): 9-11.
4. Chang Yanqiu, Xia Zixuan. Reflections on the liquidity risk monitoring and management of rural banks: Taking Baotou Municipal District as an example [J]. Northern Finance, 2021.
5. Liu Yumin. Quantitative research on the relationship between ownership structure and company performance [D]. Systems Engineering Theory and Practice, 2006.
6. Altman, E. I.. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. Journal of Finance, 1968.23 (4): 589-609.
7. Lane William R.,Looney Stephen W.,and Wansley James W..An Application of the Cox Proportional Hazards Model to Bank Failure. Journal of Banking & Finance,1986,10 (4): 0–531.
8. Chiaramonte L., Croci, E., and Poli, F..Should We Trust the Z -score? Evidence from the European Banking Industry. Global Finance Journal,2015,28: 111-131.

9. Ferriani, F., Cornacchia, W., Farroni,P., Ferrara,E.,Guarino,F.,and Pisanti,F..An Early Warning System for Less Significant Italian Banks.Questioni Di Economia E Finanza.2019.
10. Suss, J., and Treitel, H..Predicting Bank Distress in the UK with Machine Learning. Bank of England Working Papers,2019.
11. Brauning, M.,Malikkidou,D.,Scricco,G., and Scalone, S..A New Approach to Early Warning Systems for Small European Banks. ECB Working Paper Series No 2348,2019.