Economics Letters 148 (2016) 118–121

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

Random forests-based early warning system for bank failures

Katsuyuki Tanaka, Takuji Kinkyo, Shigeyuki Hamori *

Graduate School of Economics, Kobe University, 2-1, Rokkodai, Nada-Ku, Kobe 657-8501, Japan

HIGHLIGHTS

- This paper introduces a random forests-based early warning system (EWS).
- Our approach is significantly more accurate than other conventional EWSs.
- There are 730 banks in danger with assets equivalent to about 95.3 million US dollars in total.

ARTICLE INFO

Article history:
Received 21 June 2016
Accepted 19 September 2016
Available online 5 October 2016

JEL classification:
G1
C5

Keywords:
Random forests
Early warning system
Bank failure

ABSTRACT

This paper introduces a novel random forests-based early warning system for predicting bank failures. We apply this method to the analysis of bank-level financial statements, in order to find patterns that identify banks in danger of failing. The experimental results show that our method outperforms conventional methods in terms of prediction accuracy.

1. Introduction

An early warning system (EWS) that signals a country's vulnerability to financial crises can be particularly useful for developing an effective financial safety net. Following the seminal work of Frankel and Rose (1996) and Kaminsky et al. (1998), numerous studies on EWSs have sought to identify reliable early warning indicators. In particular, the global financial crisis of 2008–09 has re-boosted interest in EWSs, and has stimulated innovative work (Rose and Spiegel, 2011; Frankel and Saravelos, 2012; Shin, 2013).

The standard approach employed in the literature on EWSs is to estimate a multivariate logistic regression, in which the probability of a crisis is related to a set of explanatory variables, such as current account balance, real exchange rates, credit growth, and fiscal balance. However, there are two major deficiencies with this approach. First, explanatory variables must be pre-selected from a wide range of economic indicators based on some prior information. Second, the logistic regression does not readily allow for non-linear or threshold effects of explanatory variables. Recognizing these limitations of logistic regressions, Ghosh and Ghosh (2002) employ a method of data analysis known as the decision tree. This method uses a sequence of crisis event prediction rules based on a vector of explanatory variables. At each node of the tree, the sample is split into two sub-branches, according to the threshold value of an explanatory variable. The process is repeated until it reaches a terminal node. This method requires no pre-selection of explanatory variables and allows for non-linear effects.

In this paper, we introduce a novel approach that uses random forests to build an EWS. Random forests, which are a variant of decision trees, significantly improve classification accuracy by building a large number of trees instead of a single tree (Breiman, 2001). It circumvents an over-fitting problem by using randomly selected input variables to split each node. Moreover, it performs better with large datasets. Random forests are used in various application areas, including computer vision and bioinformatics. Random forests are popular because they are simple flexible and can be applied to a range of tasks, including classification and regression. Our approach differs from previous works because we use bank-level financial data. While many studies use macroeconomic explanatory variables to build EWSs, recent literature on the global financial crisis indicates that individual banks' financial conditions can explain differences

* Corresponding author.
E-mail address: hamori@econ.kobe-u.ac.jp (S. Hamori).

http://dx.doi.org/10.1016/j.econlet.2016.09.024
0165-1765/

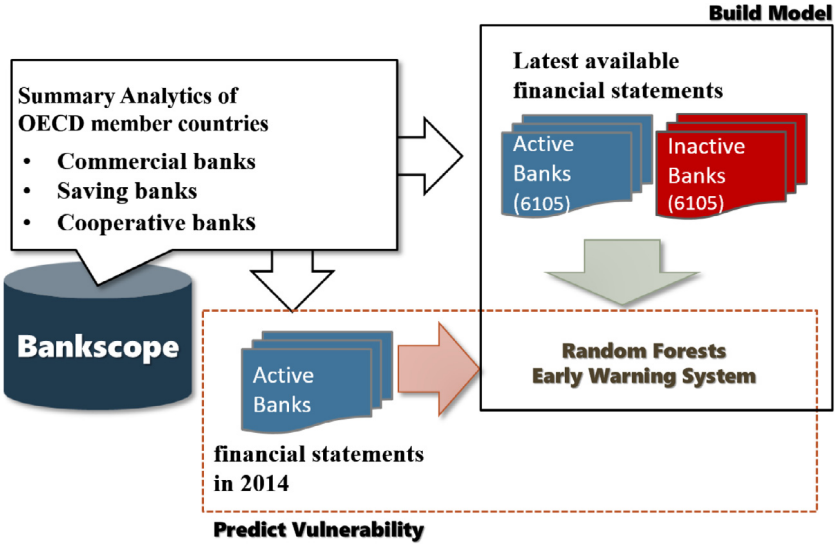

Fig. 1. Overview of random forests EWS.

in cross-bank performance observed during the crises (Berger and Bouwman, 2013; Vazquez and Federico, 2015). To predict bank failure events, we apply the random forests method to the analysis of bank-level financial data in order to identify hidden patterns that can distinguish active and inactive banks. To the best of our knowledge, this is the first paper that employs random forests to build an EWS for predicting bank failure. We call this new type of EWS the Random Forests EWS. We show that the Random Forests EWS outperforms conventional EWSs in terms of prediction accuracy.

The remainder of the paper is organized as follows: Section 2 describes data and methodology. Section 3 evaluates the performance of the Random Forests EWS and demonstrates its usefulness by assessing the vulnerability of OECD member countries based on predicted bank failure. Section 4 presents conclusions.

2. Data and methodology

2.1. Data

The data source for the bank-level financial statements is BankScope. We use 48 indicators derived from the Summary Analytics category. These indicators are classified into four groups: profitability ratio, capitalization, loan quality, and funding. We define a bank failure event as the change of a bank's status from active to inactive (i.e., bankrupt, in liquidation, or dissolved) as reported by BankScope. Our sample covers commercial banks, saving banks, and cooperatives incorporated in OECD Member countries (18,381 in total). We use the latest available financial statements of each bank; the sample period spans from 1986 to 2014.

2.2. The methodology of random forests

We apply random forests to the analysis of bank-level financial statements in order to identify patterns that distinguish between active and inactive banks, thereby producing the prediction for bank failure. More specifically, we use this method to find explanatory variables (and their threshold values) that best split the data into separate classes representing either active or inactive banks. The random forests method is a variant of decision trees with several desirable features, as discussed by Breiman (2001).

First, random forests perform better in terms of classification accuracy by building a large number of tress instead of only a single tree. Each tree is built using randomly selected data samples and randomly selected input variables from the original data to split each node. After a large number of trees are generated, they vote for the most popular class. A single-tree classifier tends to have only slightly better accuracy than a random choice of class. However, by combining a large number of trees using random input selection, the random forests can produce improved accuracy. Second, random forests provide better generalization abilities and are robust against overfitting. This is because using a random selection of input variables to split each node, as well as combining the results of multiple trees, yields error rates that compare favorably to an alternative method and are more robust with respect to noise. Third, random forests perform better with large datasets because multiple trees can be trained efficiently in parallel.

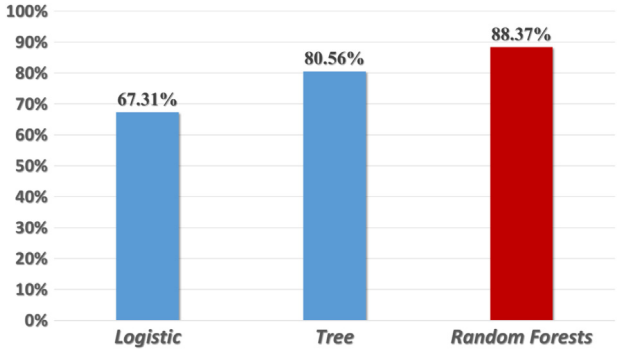

Fig. 2. Average accuracy based on ten-fold cross-validation.

Because the imbalanced size of sample data may provide misleading classification accuracy, we even out the sample sizes of active and inactive banks. As there are fewer inactive banks (6,105 banks), we use only the largest 6,105 active banks in terms of total assets to train our model. Furthermore, we eliminate variables that have missing values more than 50% of the time (there are 34 such variables). Fig. 1 illustrates the overview of the Random Forest EWS.

3. Performance evaluation and prediction

3.1. Evaluation based on ten-fold cross-validation

To evaluate the performance of the Random Forests EWS, we measure the average accuracy of classification using ten-fold cross-

Table 1
Vulnerability of OECD member countries in 2014.

	Number of banks			Total assets (Thousands of US dollars)		
	Danger banks	Country total	% share	Danger banks	Country total	% share
Austria	3	289	1.04	102	965,117	0.01
Chile	1	35	2.86	291	347,322	0.08
Germany	8	1628	0.49	853	13,458,929	0.01
Denmark	3	77	3.90	392	1,006,241	0.04
Finland	1	45	2.22	25	792,427	0.00
France	4	230	1.74	670	15,739,031	0.00
UK	3	146	2.05	464	10,317,849	0.00
Hungary	2	27	7.41	291	109,975	0.26
Iceland	1	17	5.88	44	167,741	0.03
Italy	7	525	1.33	1,824	4,026,099	0.05
Korea	2	24	8.33	344	1,340,192	0.03
Luxembourg	2	71	2.82	289	663,077	0.04
Mexico	7	65	10.77	33,152	470,784	7.04
Norway	2	130	1.54	328	596,704	0.05
New Zealand	3	24	12.50	339	343,331	0.10
Poland	2	47	4.26	395	387,798	0.10
Portugal	1	115	0.87	66	736,020	0.01
USA	730	6953	10.50	95,311	16,383,553	0.58

validation. Ten-fold cross-validation is a widely used evaluation method in the area of machine learning. It evaluates prediction accuracy by creating 10 random splits of data. In this experimental setup, nine folds are used to create a model; one fold is left out and used for testing the model Experiments are executed 10 times over 10 random splits. Using the same experimental setup, we compare the performance of the Random Forests EWS with that of conventional EWSs based on logistic regression and a decision tree. As shown in Fig. 2, the random forests method produces an accuracy rate of 88.37%, while the logistic regression and decision tree methods produce accuracy rates of 67.31% and 80.56%, respectively. The results indicate that the Random Forest EWS significantly outperforms the conventional EWSs.

3.2. Key explanatory variables

A useful property of the random forests method is that it provides importance measures for variables; this helps to identify which variables are most important for distinguishing between active and inactive banks. Hence it should provide some clues to understanding the underlying causes of bank failure. The three most important variables found in our model are as follows:

- ✓ The average rate of interest the bank is charging on its loans and other interest bearing assets. (Interest Income/Average Earning Assets).
- ✓ The average rate of interest the bank is paying on its deposits and other interest bearing liabilities. (Interest Expense/Average Interest bearing Liabilities).
- ✓ The average rate of interest the bank is charging on its loans. (Interest Income on Loans/Average Gross Loans).

3.3. Evaluation based on single-year data

We further evaluate the performance of the Random Forests EWS based on a single year of data. Using sample data from 2013, which excludes banks that became inactive before 2013, the status (active or inactive) of banks in 2013 and 2014 is predicted. By comparing the predictions with the actual status of banks, we can evaluate the prediction accuracy. Because the distribution of the sample is decidedly skewed towards active banks, we report the balanced accuracy, which is the average accuracy between the classes of active and inactive banks. In this experiment, the Random Forests EWS scored 82.48% while the logistic and the decision tree EWSs scored 52.62% and 64.60%, respectively. The results clearly indicate that the Random Forests EWS outperforms the conventional EWSs.[1]

3.4. Assessing a country's vulnerability by predicting bank failure

We use the Random Forests EWS to assess a country's vulnerability to financial crises based on predicted bank failure. Table 1 summarizes the vulnerability of each OECD member country in 2014. The vulnerability is measured by the number of banks in danger and the sum of their total assets. We define a danger bank as a bank that is predicted to change its status from active to inactive with a probability of more than 65%. According to our prediction, there are 730 danger banks with assets equivalent to about 95.3 million US dollars in total. The number of danger banks and the sum of their assets are largest in the US. However the impact of bank failure on the entire financial system could be larger in Mexico, where the shares of danger banks are 10.77% and 7.04% in terms of number and assets, respectively. Fig. A.1 in the Appendix reports the details of the predicted status of banks in all OECD member countries in 2014.

4. Conclusions

In this paper, we introduced a novel approach that uses random forests to build an EWS for predicting bank failure. Random forests are simple and flexible methods that perform better with large datasets. The results of our experiments showed that the Random Forests EWS outperformed conventional EWSs in terms of prediction accuracy. We demonstrated the usefulness of the Random Forests EWS by assessing the vulnerability of each OECD member country based on predicted bank failure. Although this paper focused on the bank failure events of advanced economies, future work might aim to explain cross-country differences in the patterns of bank failures, particularly between advanced and emerging market economies, by applying the random forests to larger datasets that contain not only bank-level data but also various macroeconomic variables.

[1] The accuracy measured for the class of active banks is 67.29%, while that for the class of inactive banks is 97.68% The former is much lower than the latter presumably because the sample is truncated by excluding smaller active banks in the former. Another possible explanation is that potentially inactive banks remained active until 2014 and they would change their status from active to inactive after 2015.

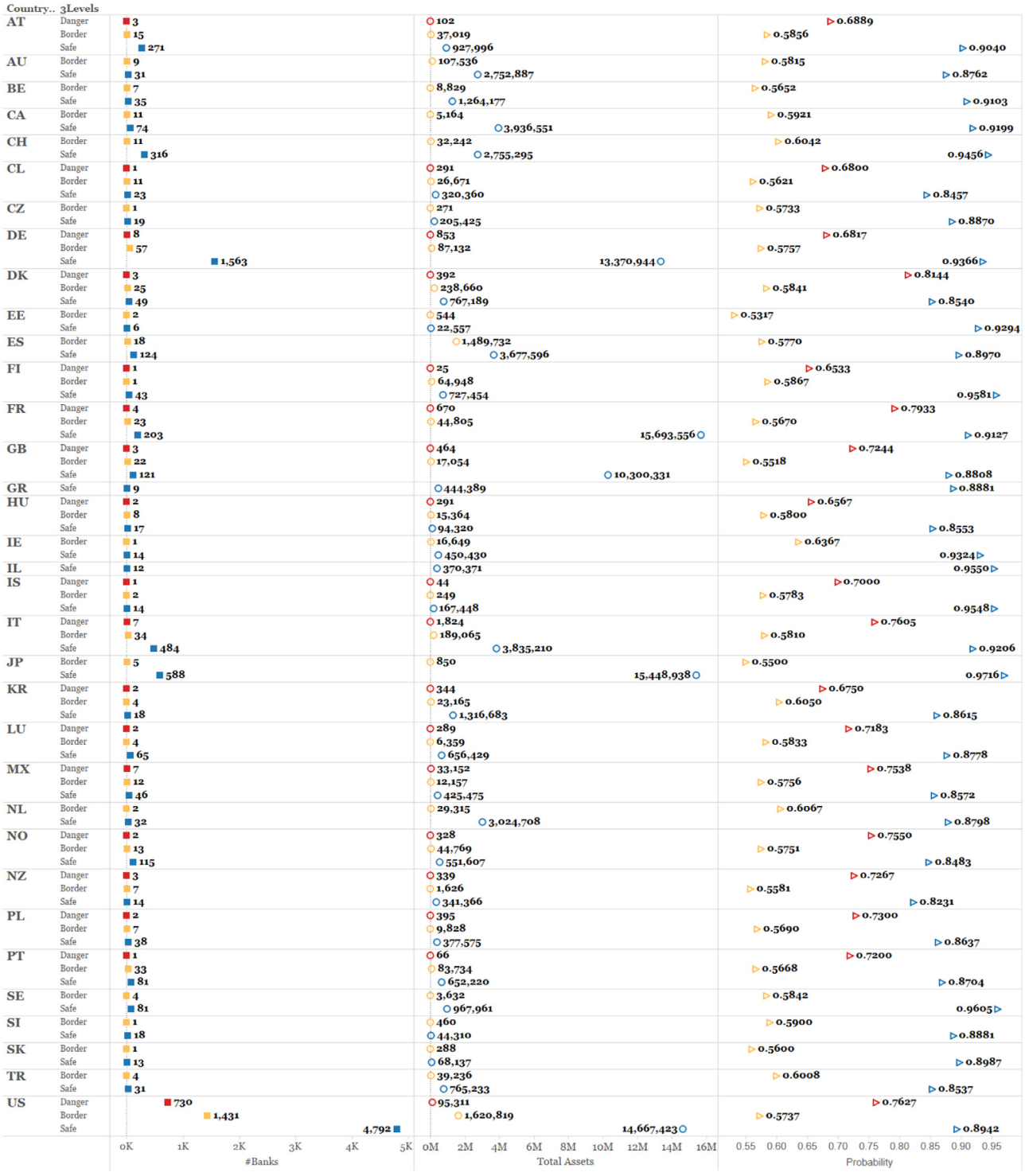

Fig. A.1. Predicted status of banks in OECD in 2014. Note: A safe bank is a bank that is predicted to remain active with a probability of more than 65%. A danger bank is a bank that is predicted to change its status from active to inactive with a probability of more than 65%. A border bank is a bank that is predicted to be either active or inactive with a probability of less than 65%.

Appendix

See Fig. A.1.

References

Berger, A.N., Bouwman, C.H.S., 2013. How does capital affect bank performance during financial crises? J. Financ. Econ. 109, 146–176.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Frankel, J.A., Rose, A.K., 1996. Currency crashes in emerging markets: An empirical treatment. J. Int. Econ. 41, 351–366.

Frankel, J., Saravelos, G., 2012. Can leading indicators assess country vulnerability? Evidence from the 2008–2009 global financial crisis. J. Int. Econ. 87, 216–231.

Ghosh, S.R., Ghosh, A.R., 2002. Structural vulnerabilities and currency crises. IMF Staff Pap. 50 (3), 481–506.

Kaminsky, G.L., Lizondo, S., Reinhart, C.M., 1998. The leading indicators of currency crises. IMF Staff Pap. 45, 1–48.

Rose, A.K., Spiegel, M.M., 2011. Cross-country causes and consequences of the crisis: An update. Eur. Econ. Rev. 55, 309–324.

Shin, H.S., 2013. Procyclicality and the search for early warning indicators. IMF Working Pap WP/13/258, International Monetary Fund.

Vazquez, F., Federico, P., 2015. Bank funding structures and risk: Evidence from the global financial crisis. J. Bank. Finance 61, 1–14.