Journal of Scientific & Industrial Research
Vol. 78, September 2019, pp. 571-575

Predicting Systemic Banking Crises using Extreme Gradient Boosting

D Alaminos[1], M A Fernández-Gámez[1], José António C Santos[2]*, and J A Campos-Soria[1]
[1]University of Malaga, 29071 Malaga, Spain
[2]Universidade do Algarve, ESGHT and CinTurs, Research Centre for Tourism, Sustainability and Well-being, 8005–139 Faro

Received 14 June 2018; revised 15 December 2018; accepted 10 May 2019

Considering the great ability of decision trees techniques to extract useful information from large databases and to handle heterogeneous variables, this paper applies Extreme Gradient Boosting for the prediction of systemic banking crises. To this end, prediction models have been constructed for different regions and the whole world. The results obtained show that Extreme Gradient Boosting overcomes the predictive power of existing models in the previous literature and provides more explanatory information on the causes that produce systemic banking crises, being the demand for deposits, the level of domestic credit and banking assets some of the most significant variables.

Keywords: Systemic banking crises; Early warning system; Extreme gradient boosting; XGBoost; Global model; Macroeconomic analysis

Introduction

In the 1990s, systemic banking crises became generalised, being the origin of the deep global economic crisis occurred in 2009, which led to research into the development of Early Warning Systems (EWS) capable of predicting these events[1,2]. Although the advances in EWS of systemic banking crises have been significant, the literature demands new models that reach greater predictability and that use broader samples, mainly including emerging and developing countries[2,3,4]. For example, Dabrowski, Conrad and de Villiers[4]used dynamic Bayesian network methods achieving 74% accuracy from a sample of European countries.Hamdaoui[5] obtained an accuracy of 84.33% using the Bayesian moving average method with data from developing countries. For its part, Ristolainen[3] applied artificial neural networks with samples from European and emerging countries, reaching accuracies close to 96%. The main objective of the present study is the construction of EWS models that serve to predict systemic banking crises with a capacity of precision superior to previous studies. Also, our work uses samples from more extensive geographic areas than those used in previous works, taking developed, emerging and developing countries. We try to fill this objective with the creation of specific models for the regions (Latin America, Asia, Africa, the Middle East, and Europe) and a global model that considers all regions of the world. For this, the present work proposes Extreme Gradient Boosting methodology (XG Boost), which is a variant of decision trees techniques. This methodology has shown high precision in previous macroeconomic analyses and supplies a more significant amount of interpretable and concise information[6].

*Author for Correspondence
E-mail: jasantos@ualg.pt

Methodology

Extreme Gradient Boosting (XGBoost)

XGBoost is an efficient and scalable gradient augmentation machine that has demonstrated a higher prediction power than other popular algorithms in recent years[7,8,9]. It is an ensemble model that is used for supervised learning problems, and that consists of classification and regression tree sets (CARTs). To predict a variable y_i, XGBoost defines the model of the form expressed in equation (1).

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) f_k \in F \quad \dots (1)$$

where K is the total number of trees, $f_k$for the tree kih is a function in the functional space F, and F is the set of all possible CARTs.

In the training phase, each of the trained new CARTs will try to complement the residual errors committed by the model. The objective function has been optimized in (t + 1) according to (2).

$$obj = \sum_{i=1}^{n} l(y_{i,} \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i) \quad \dots (2)$$

Where $l(.)$ denotes the training loss function, y_i is the value of validation in training, $\hat{y}_i^{(t)}$ is the prediction value in step t and $\Omega(f_i)$ is expressed according to the regularization term (3).

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad \ldots (3)$$

In equation (3), T represents the number of leaves, and w_j is the score on the leaf j_{th}. When optimized (2), the Taylor expansion is used so that the descent of the gradient is applied in different loss functions. During the training period of XGBoost, good features would be chosen as a node in the trees, which means that unused functions are abandoned.

Sensitivity analysis

The objective of the sensitivity analysis is to determine the relative importance of the independent variables in relation to the dependent variable[9]. Applying the Sobol' method[10], the total variance $V(Y)$ is decomposed according to equation (4).

$$V(Y) = \sum_i Vi + \sum_i \sum_{j>1} Vij + \cdots + V12 \ldots k \quad \ldots (4)$$

where $V_i = V(E(Y|X_i)$ and $V_{ij} = V(E(Y|X_i, X_j)) - V_i - V_j$.

For its part, the sensitivity indexes are determined by $S_i = V_i/V$ and $S_{ij} = V_{ij}/V$, where S_{ij} indicates the effect of interaction between two factors. The decomposition of Sobol allows the estimation of a total sensitivity index S_{Ti}, which measures the sum of all the sensitivity effects involved in the independent variables.

Sample and variables

The sample used in this study is made up of the 163 developed, emerging and developing countries that appear in the World Bank (World Bank Open Data). On the other hand, from this sample, a set of 28 independent variables selected from the previous literature was obtained[12,13,3], as shown in Table 1. These variables are macroeconomic and refer to the period 1970-2017. In turn, the dependent variable is defined by the criteria proposed by the International

Table 1 — Independent variables

Category	Code	Definition
Macroeconomic Factors	GDP per Capita	Annual real GDP per capita at current USD
	Real GDP Growth	Annual growth of real GDP
	Inflation	Rate of change in CPI
	M2 Multiplier Growth	Annual growth of M2
	M2/Reserves	Ratio of M2 to foreign exchange reserves
	REER Overall	Deviation of real effective exchange rate from 5-year rolling mean
	Gov Spending	General government final spending as % of GDP
	Real House Price	Measures the price changes of residential housing
	Unemployment	Unemployment total as % of total labour force
	Terms of Trade	Ratio of exports plus imports to GDP
	Current Account	Current account as % of GDP
Banking Sector	Domestic Credit	Ratio of domestic credit to GDP
	Credit Private Sector	Credit to private sector to GDP
	Credit Growth	Annual growth of domestic credit
	House hold Loans	Household loans at current USD
	Real Interest Rate	The bank rate meets the short- and medium-term financing needs
	Demand Deposits	Demand deposits at current USD
	Bank Assets	Ratio of bank assets to GDP
	Foreign Bank Liabilities	Foreign bank liabilities to total bank assets
	Bank Liquid Reserves	Bank liquid reserves to total bank assets
	Gov Bank Assets	Net bank claims on central government
Financial Linkages	Oil Price	Texas crude oil price at current USD
	Market Capitalization	Mark etc apitalizationto GDP
	Stock Prices	Stock market index
	Stock Prices Volatility	CBOE Volatility Index (VIX)
Governance Factors	Kaopen	Index measuring a country's degree of capital account openness
	Conflict	Dummy variable which takes on value of 1 if a country is experiencing armed conflict, and 0 otherwise
	SFI	State fragility index

Monetary Fund (IMF): (1) significant signs of financial distress in the banking system; and (2) significant banking policy intervention measures[13]. In this sense, there is financial distress when events such as important losses in the banking system and/or bank liquidations occur. On the other hand, interventions in the banking sector are considered significant if at least three of the following six measures have been used: (1) a wide liquidity support (5 percent of deposits and liabilities to non-residents); (2) gross costs of bank restructuring (at least 3 percent of GDP); (3) significant bank nationalizations; (4) significant guarantees put into effect; (5) significant purchases of assets (at least 5 percent of GDP); (6) Frozen deposits and/or closed banks. This dependent variable is denoted as 1 in the years in which the systemic banking crisis is suffered, and 0 for the opposite case.

Results

Table 2 shows the results obtained by the XGBoost for the models of each continent and the global model. We randomly selected 500 sets of variables, to which 10-fold cross-validation was applied, randomly dividing the available set of samples by 70% for training samples, and 30% for test samples. The precision of the models with training sample amounts to 98.58%, 98.32%, 99.60%, 99.82%, 99.14% for Africa & Middle East, Latin America, South & East Asia, Europe and Global, respectively. With the testing sample, the accuracy is 98.48%, 96.71%, 98.95%, 99.59%, 98.53% for models from Africa & Middle East, Latin America, South & East Asia, Europe and Global, respectively. The results of this work show that the models developed with the XGBoost achieve a prediction capacity superior to that obtained in previous studies. It is worth highlighting the case of the global model, which achieves an accuracy of 98.53%, improving the result of the model of Ristolainen[3]. Other works have obtained precisions inferior to our results, as is the case of Dabrowski *et al.*[4], where it reaches an accuracy of 74%, using dynamic Bayesian networks. It also improves the results shown by statistical methodologies, such as the study by Caggiano, Calice and Leonida[14] with Logit, which achieves a precision of 72.3% for developing countries, and the case of Hamdaoui[5] used Bayesian model averaging, which achieved a precision of 84.33% for developed countries.

For its part, Table 3 shows the sensitivity values of all the variables used in the study. The Demand Deposits variable has been significant in all models since more deposits in relation to assets make the banking system safer[11]. This result shows differences with the work of Ristolainen[3], where the variable M2/Reserves was its most significant variable in all models. Other variables belonging to the banking sector, such as Bank Assets, Domestic Credit, Credit Growth and Bank Liquid Reserves, have also been highly significant in three of the four models, showing that crisis events can occur due to liquidity problems[12], risk and return on assets problems[4]. Even so, other previous empirical works did not obtain high significance with these variables. Concerning macroeconomic variables, the most significant

Table 2 — Results of accuracy evaluation

Sample	Classification (%)		RMSE		ROC Curve	Significant Variables
	Training	Testing	Training	Testing		
Africa & Middle East	99.58	98.48	0.12	0.19	0.95	Real GDP Growth, M2 Multiplier Growth, REER Overall, Demand Deposits, Bank Liquid Reserves, Domestic Credit, Market Capitalization, Oil Price
Latin America	98.32	96.71	0.19	0.27	0.93	Real GDP Growth, Current Account, Inflation, M2/Reserves, REER Overall, Bank Assets, Credit Growth, Demand Deposits
South & East Asia	99.60	98.95	0.11	0.16	0.97	M2 Multiplier Growth, Bank Assets, Demand Deposits, Bank Liquid Reserves, Domestic Credit, Market Capitalization, Oil Price
Europe	99.82	99.59	0.08	0.12	0.98	Real GDP Growth, GDP per Capita, M2/Reserves, Demand Deposits, Bank Assets, Credit Growth, Credit Private Sector, Domestic Credit, Household Loans
Global	99.14	98.53	0.15	0.22	0.97	Inflation, M2/Reserves, REER Overall, Bank Assets, Demand Deposits, Bank Liquid Reserves, Credit Growth, Credit Private Sector, Market Capitalization

Table 3 — Variable importance values of variables for Systemic Banking Crises

Variables	Africa & Middle East	Latin America	South & East Asia	Europe	Global
GDP per Capita	0.000	0.000	0.000	0.325	0.000
Real GDP Growth	1.355	0.547	0.000	1.124	0.000
Inflation	0.000	0.362	0.000	0.087	0.142
M2 Multiplier Growth	0.526	0.000	0.142	0.000	0.000
M2/Reserves	0.000	0.425	0.000	0.572	0.525
REER Overall	0.637	1.042	0.000	0.000	0.272
Gov Spending	0.000	0.000	0.000	0.000	0.000
Real House Price	0.000	0.000	0.000	0.000	0.000
Umemployment	0.000	0.000	0.000	0.000	0.000
Terms of Trade	0.000	0.000	0.000	0.000	0.000
Current Account	0.000	0.255	0.000	0.000	0.000
Domestic Credit	0.318	0.848	0.572	0.127	0.000
Credit Private Sector	0.000	0.096	0.000	0.325	0.142
Credit Growth	0.000	0.275	0.502	0.731	1.075
Household Loans	0.000	0.000	0.000	0.089	0.000
Real Interest Rate	0.000	0.000	0.000	0.000	0.000
Demand Deposits	0.526	0.362	0.117	0.427	0.125
Bank Assets	0.000	0.425	0.970	0.572	0.502
Foreign Bank Liabilities	0.000	0.000	0.000	0.000	0.000
Bank Liquid Reserves	0.125	0.000	1.204	0.000	0.327
Gov Bank Assets	0.000	0.000	0.000	0.000	0.000
Oil Price	0.754	0.000	0.224	0.000	0.000
Market Capitalization	0.133	0.000	0.375	0.000	0.275
Stock Prices	0.000	0.000	0.000	0.000	0.000
Stock Prices Volatility	0.000	0.000	0.000	0.000	0.000
Kaopen	0.000	0.000	0.000	0.000	0.000
Conflict	0.000	0.000	0.000	0.000	0.000
SFI	0.000	0.000	0.000	0.000	0.000

variables are Real GDP Growth and M2/Reserves. They show that slow growth of GDP and a low ratio of M2/Reserves are a warning signal to predict crises. These results are consistent with those shown by other previous studies such as Ristolianen[3], Dabrowski *et al.*[4], Hamdaoui[5], but also shows differences with the results of Caggiano *et al.*[14]. On the other hand, the variable Inflation has not been highly significant, unlike other previous works such as that of Dabrowski *et al.*[4] and Ristolainen[3]. In the same line, the variable Terms of Trade has not been significant in any of the estimated models, unlike the results shown by Caggiano *et al.*[14]. Finally, none of the government variables has been significant, unlike the results obtained by Caggiano *et al.*[14].

Conclusions

The systemic banking crises are an international concern that has been a focus of attention for researchers and macroeconomic policymakers in the last decade. Our results show that decision trees, specifically the XGBoost classifier, improve the accuracy of systemic banking crises prediction models. Likewise, they offer more information for policymakers in the regions considered that require empirical tools to mitigate and resolve the negative impact of an abrupt crisis in the banking system. Also, our models can be of special relevance for financial institutions that need to control the risk of a possible imminent crisis. Given the relevance of the issue studied in this paper, the presentation of a global prediction model, with a testing accuracy of 98%, shows a significant advance in the difficult task of forecasting future adverse events on the crisis of the banking system, since the previous works have reached similar precisions only with observations of samples from small geographical areas. In addition to this high precision, it has been determined that the most significant variables that predict a systemic banking crisis are Inflation, M2/Reserves, REER Overall, Bank Assets, Demand Deposits, Bank Liquid Reserves, Credit Growth, Credit Private Sector and Market Capitalization. This model provides a unique international experience, simplifying and reducing the resources and efforts to create different models for predicting systemic banking crises.

References

1 Wilms P, Swank J, & de Haan J, Determinants of the real impact of banking crises: A review and new evidence, *N Amer J Econ Finance*, **43**(2018), 54-70.

2 Martins, AI, Efficiency determinants in Portuguese banking industry: an application through fractional regression models, *Tour Manage Studies*, **14**(2) (2018), 63-71.

3 Ristolainen R, Predicting banking crises with artificial neural networks: The role of nonlinearity and heterogeneity, *Scand J Econ,* **120**(1) (2018), 31-62.

4 Dabrowski JJ, Conrad B C,& de Villiers, JP, Systemic banking crisis early warning systems using dynamic Bayesian networks, *Expert Syst Appl*, **62** (2016), 225-242.

5 Hamdaoui M, Are systemic banking crises in developed and developing countries predictable? *J Multinat Finan Manage*, **37-38** (2016), 114-138.

6 Sevim C, Oztekin A, Bali O, Gumus S, & Guresen E, Developing an early warning system to predict currency crises, *Eur J Oper Res*, **237**(3) (2014), 1095-1104

7 Chen T, & Guestrin C, XG Boost: A scalable tree boosting system, Retrieved March 9, (2016), from the ar Xiv database (arXiv:1603.02754).

8 Lee TC, Chung PH, & Shyu JZ, Performance evaluation of medical device manufacturers using a Hybrid Fuzzy MCDM, *J Sci Ind Res*, **76**(1)(2017), 28-31.

9 Chang YC, Chang KH, & Wu GJ, Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions, *Appl Soft Comput*, **73**(2018), 914-920.

10 Groen EA, Bokkers EAM, Heijungs R,& de Boer IJM, Methods for global sensitivity analysis in life cycle assessment, *Int J LCA,* **22**(7)(2017), 1125-1137.

11 Ballester-Ripoll R, Paredes E G, & Parajola R, Sobol tensor trains for global sensitivity analysis, *Realib Eng Syst Safe*, **183** (2019), 311-322.

12 Kauko K,How to foresee banking crises? A survey of the empirical literature,*Econ Systems*, **38**(2014), 289-308.

13 LaevenL, & Valencia F, Systemic banking crises revisited, *IMF Working Paper*, (2018), WP/18/206.

14 Caggiano G, CaliceP,&Leonida L, Early warning systems and systemic banking crises in low-income countries: A multinomial logit approach, *J Bank Finance*, **47**(2014), 258-269.