Predicting of Banking Stability Using Machine Learning Technique of Random Forests

Agus Afiantara¹, Bagus Mahawan¹ and Eka Budiarto¹* 

¹ Department of Information Technology, Swiss German University, Tangerang 15143, Indonesia
* Corresponding author: eka.budiarto@sgu.ac.id

Abstract. The purpose of this research is to create a predicting model of banking stability in Indonesia. Authors use a small set of explanatory indicators primarily related to the banking industry and some relevant economic variables. Among the indicators candidate to be used in this study are the indicator of banking industry, the money markets, capital markets and creditors, and the macro-economic indicator. The source of data in this research is obtained from CEIC and SEKI (Indonesian Economic and Financial Statistics) that published by Central Bank of Indonesia from 2004 and 2011. Principal Component Analysis is used to avoid the multi-collinearity problem when construct the model. Authors train the model using Random Forest Regression with data over the 2004-2007 period, and made predictions of global financial crisis that happened in 2008/9. Python 2.7.10 and scikit-learn version 0.20.0 module has been exploited for simulations and evaluation of the model. Numerical illustration is provided to demonstrate the efficiency of proposed model. As the result nine most components analysis obtained as input for the machine learning model with explained variance ratio around 97%, accuracy around 89%, and precision 91% and mean absolute error around 11%.

1. Background
The stability of financial system is the most important priority of the government concern and its unpredictability is a critical concern of people in the world and try to avoid the events with creating a model that predict and simulate the events, especially the banking sector. The terrible events called “financial crisis” occurred in the past decade in the worlds, there are some events recorded in our memory especially in Asia and the world. In 1994-1995 there has been crisis in Mexico called “Mexican Crisis”, in 1997-1998 there has been crisis in Asia and called “Asian Crisis” (Illing and Liu, 2003), and in 2007-2009 there has been crisis called “Global Financial Crisis” (Popovska, 2014).

In Indonesia, the banking sectors still dominates Indonesia’s financial system. Based on data of BI shown that the market shares of industry reached 78.8% by the end of 2013, up from 77.9% in the first half of the same year. However, the potential risks of this industry remain to be worry of. Financial disturbance such as banking crisis in a country such as Indonesia can costly and affect to the other sectors in economic in a deeply traumatizing way. Consequently, it is important to anticipate the risks and try to prevent disaster and ensure the banking stability.

2. Literature Review
2.1. Banking Stability
Macroeconomic environment with stable condition is very important for the stability of the banking sector, mainly due to uncertainty about shaky macroeconomic and fundamental policies, such as economic growth and inflation, making it a challenge for banks to accurately assess credit risk (Kristína Kočišová, 2015). Weak economic growth, due to macroeconomic uncertainty or for other reasons, can disrupt the health of banks because it reduces the service capacity of corporate and household debt (Swamy, 2013). The resilience and stability of the financial system, especially the banking sector, as its foundation has attracted public attention even more because of the difficulties facing the global market during the crisis (Popovska, 2014).

The banking system has the important role played in propagating crisis, therefore underway to devise ways of build a strong and resilient banking system (Ghosh, 2011). The Asian crisis for more than a decade ago and even recently, after the banking crisis in Argentina and Turkey, policy makers have made proactive efforts to restructure their banking system (Hawkins and Turner, no date). The idea
behind the strategy is to maintain the banking system back to the position of profitability and solvency using private sector solutions and resolutions assisted by the public sector, or a combination thereof.

2.2. Principle Component Analysis
Principle Component Analysis (PCA) is statistical concept applied to fields such as face recognition and image compression, and common technique for finding patterns in data of high dimension. It covers mathematical concept of standard deviation, covariance, eigenvectors and eigenvalues (Smith, 2002).

PCA can be derived from a number of starting points and optimization criteria. The most important of these are minimization of the mean-square error in data compression, finding mutually orthogonal directions in the data having maximal variances, and de-correlation of the data using orthogonal transformations (Ilin, 2010).

2.3. Machine Learning and Random Forest
Machine learning is one of the technologies with ability to learn and a research field between statistics, artificial intelligence, and computer science and is also known as predictive analysis or statistical learning.

The usefulness of machine learning to predict and model has been found in the several studies. Machine learning techniques have been widely used in predicting pressure on banks and have become a major topic that is widely discussed during the pre-crisis and post-crisis periods (Ronnqvist and Sarlin, 2015). The Random Forest (RF) is one of machine learning technique used to predict bank bankruptcy in a sample of US-based financial institutions (Petropoulos et al., 2017).

The implementation of RF is widely used in many fields especially in prediction, now casting and forecasting. There are samples of implementation of RF, it has the potential to give early warning of recessions (Nyman and Ormerod, 2016), and a superior out of sample and out of time predictive performance in differing the banks of predictors and possible bank insolvencies (Petropoulos et al., 2017), insurers’ insolvency prediction (Kartasheva and Traskin, 2013), predicting the direction of stock market prices (Khaidem, Saha and Dey, 2016), and modelling of economic phenomena (David, 2017).

3. Materials and Methods
The machine learning model that generated during research activity based on figure 1. The first step is determined the crisis event represented by the value of indicator tools that called ISSK (Financial System Stability Index), it developed by Central Bank of Indonesia in 2013 as cited in (Wimanda, Maryaningsih and Nurliana, 2014).

The crisis event is classified into 4 level of stability, there are normal, alert, standby, and suspected crisis level. The ISSK indicator described in figure 2 (Wimanda, Maryaningsih and Nurliana, 2014).

![Figure 1. The process of Machine Learning Modelling](image1)

![Figure 2. ISSK Indicator Index (Image Source: Macro-prudential Policy Department of Central Bank of Indonesia (2014))](image2)
From the ISSK data extraction phase we have collected data series as dependent variable with integer value varied from 0 to 3 as table 1, that represented the stability level of financial system of Indonesia with monthly frequency.

<table>
<thead>
<tr>
<th>Idx</th>
<th>Description</th>
<th>ISSK Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Normal Condition</td>
<td>(\leq 1.25)</td>
</tr>
<tr>
<td>1</td>
<td>Alert Condition</td>
<td>(&gt;1.25) and (\leq 1.625)</td>
</tr>
<tr>
<td>2</td>
<td>Standby Condition</td>
<td>(&gt;1.625) and (\leq 1.875)</td>
</tr>
<tr>
<td>3</td>
<td>Suspected of Crisis</td>
<td>(&gt;1.875)</td>
</tr>
</tbody>
</table>

The second step, we collected the predictors data or features. The number of features that observed in this study is 33 with period over 2000-2011 and monthly frequency. All features is defined based on our previous research (Stabilitas, Perbankan and Pegrantar, 2011). The predictors consist of internal bank, money market and debt, equity market, and macro-economic indicator. The indicators that used in this study is focused area in order to predict the stability of banking system using random forest machine learning algorithm. The model described with simplified linear equation such as equation (1).

\[
y(t) = \sum_{n=1}^{n=33} \omega_n X_{n(t-1)}
\]

Where \(y(t)\) is the dependent variables represented by ISSK value or crisis event at period \(t\), \(\omega_n\) is the coefficient or weight of each independent variables, and \(X_{n(t-1)}\) is the independent variables at the previous of period \(t\). Based on that equation 2.1, the previous period of the predictors is used to predict the next month of condition of banking system purposes.

The third step, is data preparation. In this step performed N/A treatment, transformation and normalization. The firstly is N/A treatment, there are possibilities in dataset in the real-world datasets contain missing values, often encoded as blank, NaN’s or other placeholders. A basic strategy to use incomplete datasets is to discard entire rows and/or columns containing missing values. Missing data often hinder the development of robust composite indicators. Due to missing data in several predictors for period in 2000-2003, we used the basic strategy to use is discarded over period 2000-2003. Data in period over 2004-2011 is used as dataset covered in this study. The secondly, Log transformation is performed because of the data size is varies between features in dataset. It could be used to make highly skewed distribution to less skewed and the result of this step would be data pattern more interpretable and the thirdly is Normalization is required prior to any data aggregation as indicators in a dataset often have different measurement units. In this research is used the Standardization (or z-scores) as normalization method. This method is calculated by the equation (2). The result of this method is the variables would be rescaled show they would have properties of a standard normal distribution with \(\mu = 0\) and \(\delta = 1 - N (0,1)\).

\[
z = \frac{x - \mu}{\delta}
\]

The fourth step, is data validation. The purposes of data validation is performed in order to get the error rate of the machine learning model, which can be considered close to the actual error rate of the population. If the data volume is large enough to represent the population, validation is not required. Hold-out and K-Fold Cross Validation (KFCV) are the validation techniques covered in this study. In
Hold-out technique data divided into training and test data with portion 80/20 randomly, it means 80% of dataset as training data and 20% as test data (20% of data test have represented enough to the all class of banking stability level). When we used the K-Fold technique as depicted in figure 3, the dataset is divided into 1-fold with sorted period over 2004-2011. The data from period 2004 until 2007 is used as training data, and the data from period 2008 until 2011 as test data, this validation technique made the model learnt the “mini crisis” situation and tried to predict the “global crisis” situation.

The next step, is the last step before create machine learning model. PCA is the statistical procedure which used to avoid the multi-collinearity problem and reduce the dimensional of data, and finally the efficiency of resources like time, memory, and communication achieved. Fewer dimension then leads better generalization, elimination of noise to improve quality of data and further processing with machine learning algorithm (Mirkin, 2011).

The Additional step is created the machine learning model. The machine learning model is selected based on business easy of explanation and complexity. Random Forest algorithm is not too complex but not too simple also.

The last step performed in this study is evaluated the model, in this stage we built one or more models that appeared to have high quality from the data analysis perspective. The confusion matrix is a measurement to evaluated the model and known as an error matrix in machine learning, it explained the performance of classification model on a set of test data for which the true values are known.

4. Result and Discussion
In this study we explore the most important parameters of random forest and how they impact our model in term of overfitting and underfitting. The validation curve is used to validate the model described in figure 4, 5, 6 and 7 based on each parameters of random forest.
Based on these figures, the model that we made is fittest based on the rule of thumb. If training score is high and validation score is low then the model is overfitting, else if training score is low and validation score is low then the model is underfitting and else the model is fittest.

The confusion matrix is used to evaluate the model performance could be seen in table 3.1. Based on this table the model number 2 is the best model chosen with criteria number of component analysis is 9, cumulative explained variance around 97%, accuracy around 89%, precision around 91%, recall around 89% and F1-score around 89% with error criteria MAE 0.11, MSE 0.11 and RMSE 0.33.
Table 2. Evaluation of Model Performance

<table>
<thead>
<tr>
<th>Model No.</th>
<th>RF Config</th>
<th>No. of PCA</th>
<th>Cumm.Exp. Var.</th>
<th>Acc.</th>
<th>Prec.</th>
<th>Recall</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Max_Depth=2, Random_State=0,7,42</td>
<td>1</td>
<td>60.94%</td>
<td>72%</td>
<td>63%</td>
<td>72%</td>
<td>66%</td>
</tr>
<tr>
<td>2</td>
<td>Max_Depth=4, Random_State=7</td>
<td>9</td>
<td>96.91%</td>
<td>89%</td>
<td>91%</td>
<td>89%</td>
<td>88%</td>
</tr>
<tr>
<td>3</td>
<td>Max_Depth=8, Random_State=7</td>
<td>7</td>
<td>94.54%</td>
<td>89%</td>
<td>84%</td>
<td>89%</td>
<td>86%</td>
</tr>
<tr>
<td>4</td>
<td>Max_Depth=10, Random_State=0</td>
<td>7</td>
<td>94.54%</td>
<td>89%</td>
<td>84%</td>
<td>89%</td>
<td>86%</td>
</tr>
</tbody>
</table>

The confusion matrix of selected model shown in table 2. It means from 18 points of data, only 2 data missed to predict of the class of 1 (bank stability level in alert condition).

Table 3. Confusion Matrix of Selected Model

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Actual</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>13</td>
<td>3</td>
</tr>
</tbody>
</table>

The selected model (prediction) is used to predict the ISSK indicator (actual) as the reference of financial system stability index in Indonesia. The prediction model with the random split validation in figure 8, performed the regression of prediction model is earlier than actual in “mini crisis” and “global crisis”. The machine learning model is being a leading indicator for banking stability. The prediction model with k-fold cross validation (k=1) in figure 9. The predictive models have no prior detection ability, it means the amount of data training is more required.

Figure 8. Prediction vs Actual based on hold-out validation
In this last section the selected machine learning model is compared to available indicators that used by regulator of Indonesia, ISSK and BSI. Figure 10 shows that the selected model is lead earlier detection when stability of banking and financial system getting worst.

**Figure 9.** Prediction vs Actual based on 1-fold cross validation

**Figure 10.** Comparison of BSI, ISSK and selected model

### 5. Conclusions and Recommendations

The ISSK and BSI are the keys as legal indicator in Indonesia built by regulators, and as initial starting point of this study in order to build the machine learning model and proxied to actual condition of financial and banking stability of Indonesia (Wimanda et.al, 2014). Based on ISSK we classified the level of stability that indicated by integer value of 0 to 3 range of stability level and marked as dependent variable (actual value). RF algorithm could predict the actual condition of banking stability. RF algorithm should be explained and backed up with more rigorous computation and data to produce more convincing state of the art. Perhaps using other countries data for test bed.

The data differences could be used to be analysed deeper to measure dynamic processing in order to predict the stability of banking system. Beside that the knowledge of multi system and equations about how variable impact the other variable in the financial system (the transmission channel in financial system) could be nice to explored in the next study.

This methodology in this study could be tried in the next research for outside financial system in the neighbour countries of Indonesia for example, because the contagion effect of crisis should be monitored and to be an input to next indicator in order to make better prediction and study the impact of instability of financial system in other country.
Machine learning algorithm could be used and implemented to detect the stability of overall financial system especially in Indonesia. The amount of data is required and increased in order to our machine learning model more learn with huge data and the prediction should be more accurately and precisely.

**Acknowledgement**
I wish to thank to my fellows and my colleagues who want to be used as a place to discuss a lot about my thesis, there are two of my friends who give me inspirations in writing they are Hendra Syamsir and Adis Budiman. Thank you for all the information and lessons that have been given. Finally, I would like to all of my student colleagues in SGU, salute to all of you.

**References**


Early Detection of Failed Bank Through Analysis of Financial Ratios and Bank Shareholders Ratios Using Data Mining For Rural Banks

H M Agista¹, E Budiarto¹* and B Mahawan¹

¹ Business Informatics, Swiss German University, Tangerang 15143, Indonesia
* Corresponding author: eka.budiarto@sgu.ac.id

Abstract. This study aims to determine the effect of 8 bank financial ratios such as BOPO (operational efficiency ratio), CAR (Capital Adequacy Ratio), NPL (Non Performing Loan), ROA (Return On Assets), CR (Cash Ratio), KAP (quality of productive assets), PPAP (provision for loan losses) and LDR (Loan Deposit Ratio) and another ratio, namely Bank’s Shareholder ratio towards bank predictions whether a rural bank will be declared as failed bank or not. Eight financial ratios and another ratio that comparing BOD and BOC to Bank's Shareholders can be obtained from quarterly rural bank’s financial reports that have been published on the IFSA website from 2014 until 2018. The data in this research is approximately 1000 rural banks for training dataset. The method to predict rural bank become failed bank is data mining. The training dataset used is an imbalance dataset. In order to be balanced, the SMOTE method is used. The balance dataset was then analyzed with the data mining process. The data mining methods used are KNN and Naïve Bayes, both are classification method.

1. Introduction

Deposit insurance is being implemented in many countries to protect bank depositors, in full or in part, from losses caused by a bank's inability to pay its debts when due. Deposit insurance systems are one component that promotes financial stability. One of the countries that have deposit insurance corporation is in Indonesia, called IDIC (Indonesia Deposit Insurance Corporation) or Lembaga Penjamin Simpanan in Bahasa Indonesia. One of the IDIC duties is to handle a failed bank (Indonesia Deposit Insurance Corporation, 2018a).

Before becoming a failed bank, that bank would become under special surveillance. IDIC received some notices on under special surveillance bank from IFSA (Indonesia Financial Services Authority) or Otoritas Jasa Keuangan in Bahasa Indonesia. IDIC, together with IFSA, examined those banks and conduct assessment on their financial condition to assess that bank will be a failed bank or not. If become a failed bank, IDIC must handle the failed bank with several methods of resolution bank (Indonesia Deposit Insurance Corporation, 2019). It would be better if IDIC can predict which bank will be assigned as a failed bank so that IDIC can early prepare its resources and become more careful planning to deal with the bank.

To make an early detection or prediction which bank will become a failed bank, this research will use financial ratios and bank shareholder ratio. This study raises the topic of whether there is a relationship between bank financial ratios and bank shareholders to predict which bank will become a failed bank, especially for rural banks. Rural banks are the main focus because there are 91 rural banks out of the total liquidated banks until the end of 2018. (Indonesia Deposit Insurance Corporation, 2018b)

The data source that will be used is the rural bank’s publication reports published in the IFSA’s website and failed bank status for the label or class (Indonesia Financial Services Authority, 2018)(Indonesia Deposit Insurance Corporation, 2018b). The scope of the study is limited to the rural bank. The data is obtained from rural bank quarterly publication reports with restrictions from the period 2017 to 2018 and rural bank data that are stated as a failed bank. The data collected is used to create data mining models. The model formed will be used to make predictions to determine whether a rural bank will be declared as a failed bank or not. This research is different from previous research that can be explained in Table 1.