IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 10, No. 2, June 2021, pp. 407~413
ISSN: 2252-8938, DOI: 10.11591/ijai.v10.i2.pp407-413

Estimating probability of banking crises using random forest

Sri Hartini[1], Zuherman Rustam[2], Glori Stephani Saragih[3], María Jesús Segovia Vargas[4]
[1,2,3]Department of Mathematics, Universitas Indonesia, Depok, 16424, Indonesia
[4]Department of Financial Economy and Accounting I, Universidad Complutense de Madrid, Madrid, 28223, Spain

Article Info

Article history:

Received Mar 9, 2020
Revised Mar 2, 2021
Accepted Apr 16, 2021

Keywords:

Banking crises
Machine learning
Prediction of banking crises
Probability of banking crises
Random forest
Random forest regression

ABSTRACT

Banks have a crucial role in the financial system. When many banks suffer from the crisis, it can lead to financial instability. According to the impact of the crises, the banking crisis can be divided into two categories, namely systemic and non-systemic crisis. When systemic crises happen, it may cause even stable banks bankrupt. Hence, this paper proposed a random forest for estimating the probability of banking crises as prevention action. Random forest is well-known as a robust technique both in classification and regression, which is far from the intervention of outliers and overfitting. The experiments were then constructed using the financial crisis database, containing a sample of 79 countries in the period 1981-1999 (annual data). This dataset has 521 samples consisting of 164 crisis samples and 357 non-crisis cases. From the experiments, it was concluded that utilizing 90 percent of training data would deliver 0.98 accuracy, 0.92 sensitivity, 1.00 precision, and 0.96 F1-Score as the highest score than other percentages of training data. These results are also better than state-of-the-art methods used in the same dataset. Therefore, the proposed method is shown promising results to predict the probability of banking crises.

Corresponding Author:

Sri Hartini
Department of Mathematics
Universitas Indonesia
Depok, 16424, Indonesia
Email: sri.hartini@sci.ui.ac.id

1. INTRODUCTION

Banking crises are costly to the economic crisis, not only because of the high direct cost saving but also because of adverse effects on the economy [1]. Crisis in banks can significantly lessen the global economic growth by slowing down economic activities, limiting the number of stable currencies (exchange rate) for emerging economies, and weighing on their capacity to pay their debts [2]. Banking stability itself continues to receive heightened attention since the global financial crisis, circa 2008. Notwithstanding the importance of these preventative efforts, it is essential to understand the effectiveness of banking sector stability as a buffer to the real economy, when crises do occur [3]. Much of this attention has been focused on the prevention of future systemic crises.

Statistical method is a standard approach to estimate the probability of banking crisis, but low accuracy has been obtained by some standpoints of statistical techniques and the disadvantage of statistical method is we should follow the assumption of each method, if the data is not fitted to the assumptions, we cannot used the method. Then, by the devolepment of computational method, one of the methods to prevent banking crises that reaserchers usually used is machine learning. Several types of research have proved that machine learning could be a tool in predicting the banking crisis. Beutel *et al*. [4] suggested that further enhancements to machine learning early warning models are needed before they are able to offer a substantial

value-added for predicting systemic banking crises. Martinez [5] succeed in using a classification tree for predicting the financial crises on the global financial development database, which is the world bank's extensive dataset of financial system characteristics for 203 economies with several notes for improvement. Meanwhile, Christy and Arunkumar [6] proposed the implementation and comparison of several machine learning techniques including multilayer perceptron, radial basis function (RBF) network, logistic regression and deep learning to predict the financial crisis using the datasets namely German dataset, wieslaw dataset and polish dataset, which comes with the conclusion that deep learning performed better among the other methods.

Besides all the previous research, random forest usually used, considering its robustness and insensitivity to outliers and overfitting. A random forest method is a well-known approach for classification and regression problems that are especially useful, especially when working with a large number of predictor variables, or when many possible interactions exist [7]. It was also emphasized by Roy and Larocque [8] who investigated the performance of variations of random forest method, which was based on three main ideas: robustify the aggregation method, robustify the splitting criterion, and taking a robust transformation of the response.

Random forests often make accurate and robust predictions, even for very high-dimensional problem, as stated by Biau [9], with the subsampling that plays an important role in random forests so that it needs to be tuned more carefully than other parameters [10]. Random forest was used before by Rustam and Saragih [11] in predicting bank financial failures using CAMELS, which is an international rating system used by regulatory banking authorities to rate financial institutions according to capital adequacy, asset quality, management, earnings, liquidity, and sensitivity, as a predictor variables and succeed to deliver 100 percent of accuracy. Meanwhile, artificial neural network (ANN) used by Nik *et al*. [12] to build the early warning system (EWS) in order to predict the probability of financial crisis and finding the vital factor related to the prediction. This method succeeds in predicting the probability closely to its actual value; however, the method is sensitive to the existence of missing values in data.

Therefore, in this paper, the random forest was proposed as the classification and regression tool to predict and estimate the probability of banking crises using the financial crisis database formed by a sample of 79 countries in the period 1981-1999 (annual data) that has 164 crisis samples, and 357 non-crisis cases. According to the proportions between the number of crisis bank and non-crisis bank, the dataset can be identified as the imbalanced data. In such case of imbalanced data, most of the data sample belongs to one of the categories of class as majority class and the other having minimum number of data samples as minority class. Performance in this situation usually leans towards majority data size class. For handling such type of data unevenness and analyze imbalanced data traditional classifiers are not best suitable because the performance can lead to the bias which leans towards unfair classification performance [13]. However, Ahmed *et al*. [14] have experimented that random forest algorithm is the suitable machine leaning algorithm which results comparatively greater classification accuracy rate even in the imbalance data. This statement was also agreed with the study done by More *et al*. [13] that concluded the random forest as the best suitable classifier to deal with imbalanced dataset applications.

2. RESEARCH METHOD

2.1. Data

The financial crisis database, formed by a sample of 79 countries in the period 1981-1999 (annual data), is utilized in this paper. The database has 521 samples, consisting of 164 crisis samples, and 357 non-crisis cases. Gutierrez *et al*. [15] also previously used this dataset to predict the bank crises. Eleven features described each sample, which labeled as the crises or not. The target class equals one if the bank crises happen as explain by Caprio *et al*. [16] and equals zero if the bank is not in a crisis condition. The information on crises is cross-checked with the research by Domac and Martinez-Peria [17] and with international monetary fund staff reports and financial news. Meanwhile, the features are given in Table 1.

The first feature is the monetary policy strategies, consisting of exchange rate target and monetary policy target, which are dummies. The exchange rate target has the values between zero to three; meanwhile the monetary policy target equals one during periods in which targets were based on monetary aggregates, two when the objective was increased, three when the two variables are into the objective function and zero in other cases, according to the chronology of the Bank of England survey of monetary frameworks, as stated in [18]. The second feature is the central bank independence that goes from 0 (least independent) to 1 (most independent), where the index of independence is assumed to be constant throughout each decade.

The third feature is inflation, which is the percentage change in the GDP deflator. The fourth feature is the real interest rate, the nominal interest rate minus inflation in the same period, calculated as the percentage change in the GDP deflator. Then, the fifth feature is the net capital flows to GDP. It is the capital

account plus financial account plus net errors and omissions. The sixth and seventh feature is the real GDP per capita in 1995 US dollars and the real GDP growth. The real GDP per capita in 1995 is expressed in US dollars instead of purchasing power parity (PPP) for data available only for two points in time, while the real GDP growth is defined as the percentage change in GDP Volume.

Table 1. The feature of financial crisis database

No	Feature
1	Monetary policy strategies
2	Central bank independence
3	Inflation
4	Real interest rate
5	Net capital flows to GDP
6	Real GDP per capita
7	Real GDP growth
8	Domestic Credit growth
9	Bank Cash to total assets
10	Bank foreign liabilities to foreign assets
11	Previous crises

Domestic credit growth is the eight feature that explained the percentage change in domestic credit, claims on the private sector. Bank cash to total assets as the ninth feature is the reserves of deposit money banks divided by total assets of deposit money banks. The tenth feature is the foreign bank liabilities to foreign assets, which are described as the amount of deposit money banks foreign liabilities to foreign assets. As the last feature, the previous crisis has four possibilities of value. It equals zero if the country has not had a previous crisis; one, if the country has suffered one previous crisis; two, in case of two or three previous crisis; and, three, for other cases.

2.2. Proposed method

In this paper, the random forest method, which was first introduced by Breiman [19], will be used. Random forests are a type of ensemble technique that makes predictions by averaging over the predictions of several independent base models [20]. The algorithm of random forest is given in Figure 1.

Bagging (a contraction of bootstrap-aggregating) is a general aggregation scheme, which generates bootstrap samples from the original data set, constructs a predictor from each sample, and decides by averaging [21]. It is one of the most effective computationally intensive procedures to improve on unstable estimates, especially for large, high-dimensional data sets, where finding a suitable model in one step is impossible because of the complexity and scale of the problem [22]. Besides, in this paper, each tree of the random forest is built based on the bootstrap sample, shown in Figure 2, which was drawn randomly from the original dataset using the classification and regression tree (CART) method and the entropy as the splitting criterion. The entropy has a role in controlling how each decision tree decides to split the data.

Algorithm of Random Forest for Classification.

1. Given a training set with n observations and m features.
2. For $r = 1$ to R
   a. Draw a bootstrap set with n observations from the training set.
   b. Build a decision tree from each of the bootstrap sets, where each node of the tree is randomly chosen.
   c. Fully grow each decision tree without pruning
3. Output the ensemble of decision trees.

Figure 1. Algorithm of the random forest for classification [9]

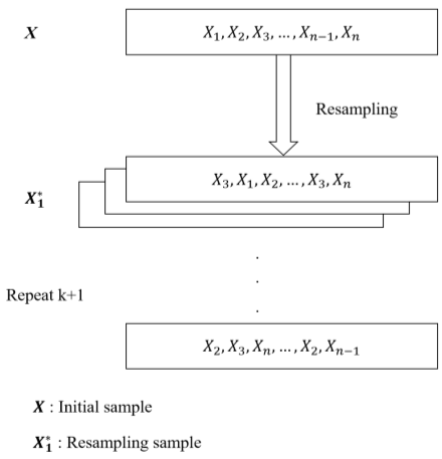

Figure 2. Schematic diagram of bootstrap resampling method [23]

In addition, random forest has several other desirable features related to low bias and low variance. First, it only requires three parameters, namely the number of trees, the number of candidates split variable at each split, and the minimum node size. This parameter usually are very easy to tune considering there are many research that suggests their recommandations values to use for obtaining optimal results [23]. Random forest can also generate an out-of-bag error, a nice estimate of the generalization error, in its growing procedure, while other models generally require multiple training procedures like cross-validation to generate such estimates. This algorithm can also generate variable importance indices in its growing procedure and turn them out to be great estimator for variable relevancies [24]. Furthermore, the random forest is a robust method against irrelevant features and outliers in training data. It has structured as a treewhich leads to expand itself in nature easy to fit more data by growing more branches. It leads the random forest algorithm becomes a very adaptive machine learning model.

As the general rule, random forest classification algorithm is used in two phases. First, the random forest algorithm extracts subsamples from the original samples using the bootstrap resampling method and creates decision trees for each sample. Second, the algorithm classifies the decision trees and implements a simple vote, with the largest vote of the classification as the final result of the prediction. The random algorithm that given in Figure 1 was also always includes three steps is: [25]

1. Select the training set. Use the bootstrap random sampling method to retrieve bunch of training sets from the original dataset with m features, with the size of each training set the same as that of the original training set.
2. Build the random forest model. Create a classification regression tree for each of the bootstrap training sets to produce some number of decision trees to form a forest. The growth of each tree was built by the approach that does not choose the best features as internal nodes for branches but rather a random selection of features of all features.
3. Create a voting. Since the training process of each decision tree is independent, the training of the random forests can work in parallel, which significantly improves efficiency. The random forest can be created by combining a number of decision trees trained in the same way. When classifying the input samples, the results depend on tha majority vote from the output of each decision tree. The random forest algorithm determines the samples by constructing a series of independent and distributed decision trees and determines the final category of the sample according to each decision tree.

For the regression purpose in determining the probability of banking crises in this paper, the number of predictions of each class divide by the number of trees. In this paper, a hundred trees are used. Therefore, if there are 60 trees that give a prediction that the bank will be in crisis while the rest of the trees provide the opposite prediction, then the probability of bank will be in crisis is 0.6.

2.3. Performance measurement

The proposed method in this paper will be evaluated using the confusion matrix, as shown in Table 2, to determine whether the method proposed can be used for estimating the probability of bank crises. The contents of Table 2 can be explained: TP is the number of crisis bank samples which were correctly predicted, FP is the total of non-crisis bank samples which were incorrectly predicted, TN is the count of non-crisis bank samples which were correctly predicted, and FN is the number of crisis bank samples which were incorrectly predicted.

Table 2. The confusion matrix

		Prediction Class	
		Crisis	Non-Crisis
Actual Class	Crisis	TP	FN
	Non-Crisis	FP	TN

An evaluation was then carried out according to the confusion matrix to calculate the accuracy, sensitivity, precision, and F1-Score of the proposed methods. According to Kotu and Deshpande [26], accuracy is the aggregate measure of classifier performance, which the formula is given in (1).

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FN+FP}} \quad (1)$$

The sensitivity is defined as the proportion of all relevant cases that were found where the formula is shown in (2).

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \quad (2)$$

Then, the precision, given in (3) is defined as the proportion of cases found that was relevant.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (3)$$

Meanwhile, considering the sensitivity and precision, F1-Score takes the formula as (4).

$$\text{F1 Score} = \frac{2 * \text{sensitivity} * \text{precision}}{\text{sensitivity} + \text{precision}} \quad (4)$$

3. RESULTS AND ANALYSIS

As explained before, the performance of random forest in classifying whether a bank will be in crisis condition or not will be evaluated according to its accuracy, sensitivity, precision, and F1-Score. Its performance was given in Table 3. From this table, we can conclude that the average of each aspect of performance measurement is 0.96 accuracy, 0.89 specificity, 0.88 precision, and 0.88 F1-Score. Besides, the performance when using 90 percent of training data is better than when using other percentages with 0.98 accuracy, 0.92 specificity, 1.00 precision, and 0.96 F1-Score.

Therefore, assumed we used the model built using 90 percent of training data. If we were given a hundred banks to be predicted, there is a 98 percent probability that we will predict it correctly. For the crisis bank cases, there is an 92 percent probability that we will correctly predict the crisis bank among all of the crisis banks that we predicted, and there is an 100 percent probability that we will correctly predict among the bank that actually crises. As the comparison between the accuracy of training and testing set using the random forest, Figure 3 gives an illustration of how the training set succeeds in reaching 100 percent accuracy for every proportion of training and testing set, while testing set seems to have up and down performance depending to its proportion.

Table 3. The performance of random forest in classifying banking crises

Percentage of training data (%)	Accuracy	Specificity	Precision	F1-Score
10	0.97	0.78	0.84	0.81
20	0.95	0.75	0.84	0.79
30	0.94	0.86	0.79	0.82
40	0.96	0.92	0.88	0.90
50	0.95	0.88	0.93	0.90
60	0.96	0.93	0.89	0.91
70	0.96	0.97	0.86	0.91
80	0.97	1.00	0.88	0.94
90	0.98	0.92	1.00	0.96

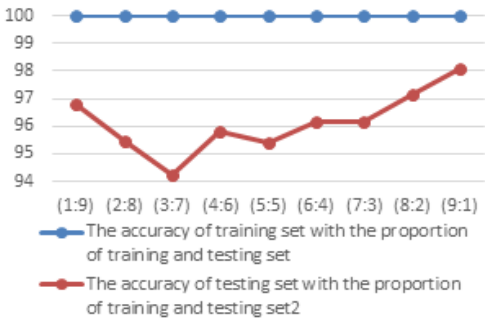

Figure 3. The comparison between the accuracy of training and testing set using the random forest model

Parallel with Table 3, we can see from Figure 3 that 90 percent of training data gives the highest accuracy, precision, and F1-Score among the others. This performance is also an improvement from the previous research done by Gutierrez *et al.* [15] which reported SimpleLogistic regression using initial

covariates and product units (LRIPU*) with the highest accuracy (91.14 percent) and Logistic Model Tree (LMT) with the highest sensitivity (88.89 percent).

Therefore, the confusion matrix from the performance of training and testing set in a random forest model was given in Tables 4 and 5, respectively. Both from those tables, the number of non-crisis banks is higher than the number of crisis banks. Furthermore, from Table 4, it shows that there is no sample that incorrectly predicted both for the crisis bank and non-crisis bank. Meanwhile, Table 5 shows that there is a crisis bank that was incorrectly predicted as the non-crisis bank.

For more detail view, Table 6 shows the probability of each class prediction. If the probability that it is a crisis bank is greater than the probability that it is a non-crisis bank, then the sample concluded as the crisis bank. As an example, a sample number 5 in Table 6 has a 0.97 probability that it is a crisis bank and has a 0.03 probability that it is a non-crisis bank, then the sample number 5 was predicted as the crisis bank. The same rule was also applied to other samples. For sample number 50, it has 0.51 probability that it is a non-crisis bank, which is higher than 0.49. Therefore, the prediction class for the sample is a non-crisis bank, which is a false prediction according to its actual class.

Table 4. The confusion matrix from the performance of training set in a random forest model

		Prediction Class	
		Crisis	Non-Crisis
Actual Class	Crisis	152	0
	Non-Crisis	0	316

Table 5. The confusion matrix from the performance of testing set in a random forest model

		Prediction Class	
		Crisis	Non-Crisis
Actual Class	Crisis	11	1
	Non-Crisis	0	41

Table 6. The prediction probability of banking crisis of testing set in a random forest model

No	The probability that a bank is in a crisis condition	The probability that a bank is not in a crisis condition	Actual Class	Prediction Class
1	0.25	0.75	0	0
2	0.11	0.89	0	0
3	0.01	0.99	0	0
4	0.10	0.90	0	0
5	0.97	0.03	1	1
…	…	…	…	…
49	0.28	0.72	0	0
50	0.49	0.51	1	0
51	0.13	0.87	0	0
52	0.10	0.90	0	0
53	0.16	0.84	0	0

4. CONCLUSION

Banking stability is essential to maintain because of its strong relationship with financial stability. Therefore, the banking crises were also crucial to prevent future systemic crises. In this paper, a random forest algorithm was proposed to classify the crisis and non-crisis bank, complete with its probability. As the dataset, Financial Crisis Database containing a sample of 79 countries in the period 1981–1999 was used with several numbers of the percentage of training data. From the experiments, the accuracy of our proposed method succeeds in providing up to 90 percent accuracy, sensitivity, precision, and F1-Score. Therefore, it can be concluded that the random forest can predict the probability of crisis or non-crisis banks accurately. As future work, an updated dataset can be utilized to analyze our proposed method for obtaining new insights. Other methods, especially the ensemble method, are also recommended to be developed.

ACKNOWLEDGEMENT

This research was supported financially by the Indonesia Deposit Insurance Corporation research grant scheme.

REFERENCES

[1] B. Bojinov, “Causes of banking crises in modern world.” in *SSRN Electronic Journal*, 2014. https://dx.doi.org/10.2139/ssrn.2438182.

[2] B. Firtescu, "Causes and effects of crises on financial system stability in emerging countries," in *Procedia Economics and Finance*, vol. 3, pp. 489-495, 2012. https://doi.org/10.1016/S2212-5671(12)00185-2.

[3] R. Stewart, and M. Chowdhury, "Does bank stability promote economic resilience?: Evidence From Panel Data," in *SSRN Electronic Journal*, 2019. https://dx.doi.org/10.2139/ssrn.3339563.
[4] J. Beutel, S. List, G. von Schweinitz, "An evaluation of early warning models for systemic banking crises: Does machine learning help us predict banking crises?" in *Journal of Financial Stability*, vol. 45, pp. 100693, 2019.
[5] N. Martinez, "Predicting financial crises." in *Wharton Research Scholars*, vol. 136, 2016.
[6] S.A. Christy, and R. Arunkumar, "Machine learning based classification models for financial crisis prediction," in *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, pp. 4487-4893, 2019. https://www.researchgate.net/deref/http%3A%2F%2Fdx.doi.org%2F10.1109%2FTSMCC.2011.2170420.
[7] A. Sage, "Random forest robustness, variable importance, and tree aggregation," in *Graduate Theses and Dissertations*, 2018. https://doi.org/10.31274/ETD-180810-6083.
[8] M-H. Roy, and D. Larocque, "Robustness of random forests for regression," in *Journal of Nonparametric Statistics*, vol. 24, pp. 993-1006, 2012. https://doi.org/10.1080/10485252.2012.715161.
[9] G. Biau, "Analysis of a random forests model," in *Journal of Machine Learning Research (JMLR)*, vol. 13, pp. 1063-1095, 2012.
[10] C. Tang, D. Garreau, and U. von Luxburg, "When do random forests fail?," in *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, Montreal, Canada, 2018..
[11] Z. Rustam, and G.S. Saragih, "Predicting bank financial failures using random forest," in *2018 International Workshop on Big Data and Information Security, IWBIS 2018* (pp. 81-86), 2018. https://doi.org/10.1109/IWBIS.2018.8471718.
[12] P.A. Nik, MansourJusoh, A.H. Shaari, and T. Sarmdi, "Predicting the probability of financial crisis in emerging countries using an early warning system: Artificial Neural Network," in *Journal of Economic Cooperation and Development*, vol. 37, pp. 25-40. 2016.
[13] A.S. More, D.P. Rana, and I. Agarwal, "Random forest classifier approach for imbalanced big data classification for smart city application domains," in *International Journal of Computational Intelligence & IoT*, vol. 1, 2018.
[14] L. Ahmed, V. Georgiev, M. Capuccini, M. *et al*. "Efficient iterative virtual screening with Apache Spark and conformal prediction," in *Journal of Cheminformatics*, vol. 10, 2018.
[15] P.A. Gutierrez, M.J. Segovia-Vargas, S. Salcedo-Sanz, C. Hervas-Martınez, A. Sanchis, J.A. Portilla-Figueras, and F. Fernandez-Navarro, "Hybridizing logistic regression with product unit and RBF networks for accurate detection and prediction of banking crises," in *Omega*, vol. 38, pp. 333-344, 2010. https://doi.org/10.1016/j.omega.2009.11.001.
[16] G. Caprio, and D. Klingebiel, "Episodes of systemic and borderline financial crises," in *Dataset mimeo, The World Bank*, 2003.
[17] I. Domac, and M.S. Martinez-Peria, "Banking crises and exchange rate regimes: is there a link?," in *The World Bank*, vol. 61, no. 1, pp. 41-72, working paper number 2489; 2000. https://doi.org/10.1016/S0022-1996(02)00081-8.
[18] L. Mahadeva, and G. Sterne, *editors*, "Monetary policy frameworks in a global context," London: Routledge, 2000.
[19] L. Breiman, "Random forests," in *Machine Learning*, vol. 45, pp. 5-32, 2001.
[20] M. Denil, D. Matheson, N. de Freitas, "Narrowing the gap: Random forests in theory and in practice," in *Proceedings of the 31 st International Conference on Machine Learning, Beijing, China*, vol. 32, 2014.
[21] A. Kleiner, A. Talwalkar, P. Sarkar, and M.I. Jordan, "A scalable bootstrap for massive data," in *arXiv*, arXiv:1112.5016, 2012.
[22] S. Wager, T. Hastie, and B. Efron, "Standard errors for bagged predictors and random forests," in *arXiv*, arXiv:1311.4555, 2013.
[23] J. Mei, D. He, R. Harley, T. Habetler, and G. Qu, "A random forest method for real-time price forecasting in Ney York electrivity market," in *IEEE Power and Energy Society General Meeting*, pp. 1-5, 2014. https://www.researchgate.net/deref/http%3A%2F%2Fdx.doi.org%2F10.1109%2FPESGM.2014.6939932.
[24] T.K. Ho, "The random subspace method for constructing decision forests," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no, 8, pp. 832-844, 1998. https://doi.org/10.1109/34.709601.
[25] X. Gao, J. Wen, and C. Zhang, "An improved random forest algorithm for predicting employee turnover," in *Hindawi: Mathematical Problems in Engineering*, vol. 2018, pp. 1-12, 2018. https://doi.org/10.1155/2019/4140707.
[26] V. Kotu, and B. Deshpande, "Model evaluation data science," (Second Edition) eds Kotu V and Deshpande B (Cambridge: Morgan Kaufmann) chapter 8 pp 263-279, 2019.