Prediction Insolvency of Insurance Companies Using Random Forest

Z Rustam and G Saragih

Department of Mathematics, Faculty of Mathematics and Science, Universitas Indonesia, Depok, Indonesia.

*e-mail: rustam@ui.ac.id

Abstract. Insurance companies have an important role in economy because they support every insured companies, so it is guaranteed financially, therefore is important to predict the insolvency of insurance companies. Insolvency prediction is like an early warning system for insurance companies. Rustam and Yaurita had predicted insolvency of insurance company by using Spanish non-life insurance firm data from Prof. Dr. Maria Jesus Segovia using support vector machines and fuzzy kernel c-means as classifiers (2018). Furthermore, this research is novel based on the use of random forest as a classifier. The result obtained is reported in percentage of accuracy, both in training and testing of random forest, which was 100% while used 50% training data with entropy and the number of estimators is 100 as hyperparameters. This classification, therefore, shows the best value in contrast with prior methods, even though only 50% of training data set was used.

Keyword: Prediction insolvency, random forest

1. Introduction

The first Insurance companies have an important role in maintaining economic stability in a country because they are large investors in financial markets. The insurance sector has been regarded as a stable segment of financial system because most insurers’ balance sheets are composed of relatively illiquid liabilities that protect them from the risk of rapid liquidity shortage that do confront most of banks [1]. Another important role for insurance companies is they protect the insured one from various risks that face them in the financial market [2]. So by holding an important role in the financial market, it is also important to take care of them so they don’t face insolvency. The insurance sector can, however, be a source of vulnerability for the financial system, and the failure or insolvency of an insurer can cause a financial instability in a country.

There were several methods have been proposed to predict failure in insurance sector, most of them are statistical methods, such as discriminant or logit analysis [3] [4] and used financial ratios as variables. However those kind of variables usually do not follow the assumptions of statistic [5]. Sergio-Vargas, et al used rough set theory to avoid the inconveniences of statistical methods [2], then with the same dataset, Rustam and Yaurita used support vector machine and fuzzy kernel c-means to increase the accuracy of predicting insololvency of insurance companies [6]. So in order to enhance the predictability, in this research we propose random forest to predict insolvency of insurance companies based on data that is also used by Sergio-Vargas, et al [5] and Rustam and Yaurita’s [6] papers. Random forest has become a commonly used tool for multiple prediction scenarios with the high
accuracy of prediction result [7]. There were several papers used random forest as a classifier, such as: classify solvent and insolvent of bank, with accuracy of 93% [8], automated diagnosis of heart disease at 83.6%, applying the weighted random forest [9] and diabetes mellitus at 80.8% [10]. Hence, those papers have proven the capability of random forest to apply to any problem, produced good model performance in the process. This research is organized as follows: Section 1 provides background of details, while Section 2 specifies data and research method used. Then, Section 3 discussed result and analysis, and finally, conclusions are included in Section 4.

2. Data and Research Method

2.1. Data
Data in this research is provided by Prof. Dr. Maria Jesus Segovia that collected sample of Spanish firms used by Sanchis that consist of non-life insurance firm data one years prior to failure [1]. This dataset has 72 firms selected, 36 failed firms and 36 healthy firms. There are 21 ratios used as variables to classify the solvent and insolvent insurance firms that calculated from the last financial statements issued before firm bankruptcy [1]. All ratios were in discrete type of input, as proposed method of Rustam and Yaurita who recorded the continuous variable of ratios into qualitative terms (low = 1; medium = 2; high = 3; very high = 4), which gave the significant effect into classifiers [2]. Table 1 shows all ratios [1].

| Ratio | Definitions |
|-------|-------------|
| R1 | Working capital / total assets |
| R2 | Earnings before taxes / (capital + reserves) |
| R3 | Investment income / investments |
| R4 | (Earnings before taxes + reserves for depreciations + provisions + (extraordinary income – extraordinary charges)) / total liabilities |
| R5 | Earned premiums / (capital + reserves) |
| R6 | Earned premiums / (capital + reserves + technical provisions) |
| R7 | Earned premiums net of reinsurance / (capital + reserves) |
| R8 | Earned premiums net of reinsurance / (capital + reserves + technical provisions) |
| R9 | (Capital + reserves) / total liabilities |
| R10 | Technical provisions / (capital + reserves) |
| R11 | Claims incurred / (capital + reserves) |
| R12 | Claims incurred / (capital + reserves + technical provisions) |
| R13 | Claims incurred net of reinsurance / (capital + reserves) |
| R14 | Claims incurred net of reinsurance / (capital + reserves + technical provisions) |
| R15 | (Claims incurred / earned premiums) + (other charges and commissions / other income) |
| R16 | (Claims incurred net of reinsurance / earned premiums net of reinsurance) + (other charges and commissions / other income) |
| R17 | (Claims incurred + other charges and commissions) / earned premiums |
| R18 | (Claims incurred net of reinsurance + other charges and commissions) / earned premiums net of reinsurance |
| R19 | Technical provisions of assigned reinsurance / technical provisions |
| R20 | Claims incurred / earned premiums |
| R21 | Claims incurred net of reinsurance / earned premiums net of reinsurance |
2.2. Research Method

2.2.1. Random Forest. Random Forest was proposed by Breiman [11]. Random forest is an ensemble machine learning method with a supervised learning algorithm that builds a large number of uncorrelated decision trees based on averaging random selection of predictor variables, this approach will reduce the variance [12]. Each decision tree will provide a classification for input data then the result will choose based on the most voted prediction [13]. The input data in each of tree is sampled data from the original dataset that is obtained by implementation of bootstrap method. Bootstrap is a resampling method that introduced by Jackknife, this sampling was performed with replacement, which there is a probability of same observations appearing more than once [14]. When constructing a decision tree, each time a split in the tree is considered and a random selection of \( m \) predictors is chosen as a subset of split candidates from the full set of predictors [12]. Hence, as the new selection of \( m \) predictors is generated at each split, and one typically chooses \( m = \sqrt{p} \), means the number of predictors considered at each split \( (m) \) is approximately equal to the square root of the total number of predictors, \( p \) [12]. There are two types of criterion that can used while split the tree, those are: entropy and gini. Random forest has ability to improve accuracy of model by randomization and voting methods, and also able to reduce the correlation between tress [15], however many trees, because Breiman has proved that the process building of numerous trees does not create an overfit, although it produces a generalization error that converges to a value [11].

Algorithm random forest [8]

1. Given the training data set, with \( n \) and \( p \) as observations and variables, respectively
2. For \( b = 1 \) to \( B \)
   a. Draw a bootstrap sample with \( n \) number of observations from the training (original) data set
   b. Build the decision tree \( T^b \) from each new result derived, where individual nodes are chosen at random.
      i. Select \( m \) variable at random from \( p \), with \( m \leq p \), where \( m = 1, 2, \ldots \) or less than \( \sqrt{p} \) [12].
      ii. Choose the best feature that provides satisfactory Information Gain or Gini Index.
      iii. Split the node
   c. Grow each without pruning
3. Output the ensemble of decision trees \( \{T^b\}_1^B \)
4. Conduct voting, i.e., if \( \hat{C}_b(x) \) is the class prediction of the \( b \)th random forest tree, then
   \[
   \hat{C}^B_{\text{rf}}(x) = \text{majority vote} \{\hat{C}_b(x)\}_1^B
   \]

2.2.2. Evaluation of Model Performance. Confusion matrix is used to help determine the performance of model, it shows the number of correct and incorrect prediction that made by the model compared to the actual data. Table 2 shows confusion matrix that used in this research.

| Actual Class | Solvent | Insolvent |
|--------------|---------|-----------|
| **Predicted Class** | **TP** | **FN** |
| Solvent      |         |           |
| Insolvent    |         |           |

Table 2. Confusion Matrix
However, the definition of 4 parts to this confusion matrix:
1. TP: True Positive, observed on instances where solvent is detected as solvent
2. TN: True Negative, is when insolvent is detected as insolvent
3. FP: False Positive, is seen in cases where insolvent is detected as solvent
4. FN: False Negative, is when solvent is detected as insolvent

Based on the value of TP, TN, FP, and FN from the matrix, it is possible to obtain the valued accuracy with the following formula:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

3. Results and Analysis
A study [2] applied support vector machines and fuzzy kernel c-means on the same non-life insurance data, therefore, the aim in this research is a novelty by using random forest to increase predictability.

Hyperparameters tuning is the way to obtain the best model that affect to the model structure and also the result of the output [16]. The hyperparameters tuning always depends on the data set, but tune hyperparameters is recommended, if no tuning is conducted, it cannot be ensured that the model is the best predictive model for that dataset [17]. This research was tuning 2 hyperparameters, those were criterion for impurity (entropy and gini), and number of estimators (tree), so we should have several models from hyperparameters combinations. Then, this research required that the algorithm was run 10 times, and the repetition was performed due to the presence of random in this experiment.

Table 3 shows the accuracy for predicting the insolvency of insurance by using random forest with entropy as a criterion, while gini was a criterion in Table 4.

| Percentage of Data Training | Number of Trees | 
|-----------------------------|-----------------| 
|                             | 10 | 50 | 100 | 10 | 50 | 100 |
| 10%                         | 83% | 85% | 96% | 100% | 100% | 100% |
| 20%                         | 83% | 89% | 93% | 100% | 100% | 100% |
| 30%                         | 80% | 92% | 100% | 100% | 100% | 100% |
| 40%                         | 86% | 93% | 97% | 100% | 100% | 100% |
| 50%                         | 91% | 92% | 100% | 100% | 100% | 100% |
| 60%                         | 93% | 96% | 100% | 100% | 100% | 100% |
| 70%                         | 100% | 100% | 100% | 100% | 100% | 100% |
| 80%                         | 100% | 100% | 100% | 100% | 100% | 100% |

Based on Table 3, random forest with entropy was able to correctly predict the insolvency of insurance with 100% accuracy for all compositions of training data set and number of trees. For testing data, the highest accuracy was obtained while the percentage of data training was 30%, 50%-80% with number of trees 100, that is 100%.
Table 4. Accuracy of Prediction Insolvency of Insurance Using Random Forest with Entropy as Criterion

| Percentage of Data Training | Number of Trees | Accuracy of Testing Data | Accuracy of Training Data |
|-----------------------------|----------------|--------------------------|---------------------------|
|                             | 10  | 50  | 100 | 10  | 50  | 100 |
| 10%                         | 72% | 80% | 84% | 100%| 100%| 100%|
| 20%                         | 86% | 87% | 94% | 100%| 100%| 100%|
| 30%                         | 80% | 88% | 92% | 100%| 100%| 100%|
| 40%                         | 86% | 88% | 93% | 100%| 100%| 100%|
| 50%                         | 83% | 88% | 97% | 100%| 100%| 100%|
| 60%                         | 89% | 93% | 96% | 100%| 100%| 100%|
| 70%                         | 100%| 100%| 100%| 100%| 100%| 100%|
| 80%                         | 100%| 100%| 100%| 100%| 100%| 100%|

Based on Table 4, random forest with gini was able to correctly predict the insolvency of insurance with 100% accuracy for all compositions of training data set and number of trees, as well as entropy result. For testing data, the highest accuracy was obtained while the percentage of data training was 70% – 80% with all composition of number of trees, showed that the accuracy of each model is 100%.

Based on Table 3 and Table 4, it is seen that a greater composition of tree that used is able to produce higher accuracy, due to the presence of more data learned by the model. Hence, tuning indicates the best accuracy on instances where the criterion entropy is with number of estimators as 100. However, it is seen that the results are similar for each composition training data and number of estimators.

As said in Section 2, Insurance firm data was followed that same way as in the paper written by Rustam and Yaurita (2018), due to a desire to compare performance results (accuracy) of the different methods used. Table 5, therefore, shows the comparison of each methods’ Testing accuracy.

Table 5. Performance Results

| Training Data (%) | SVM (%) | Fuzzy Kernel C-Means (%) | Random Forest (%) |
|-------------------|---------|--------------------------|-------------------|
| 10                | 70.31%  | 56.25%                   | 96%               |
| 20                | 71.43%  | 58.93%                   | 93%               |
| 30                | 68%     | 60%                      | 100%              |
| 40                | 66.67%  | 61.9%                    | 97%               |
| 50                | 66.67%  | 66.67%                   | 100%              |
| 60                | 67.86%  | 71.43%                   | 100%              |
| 70                | 65%     | 75%                      | 100%              |
| 80                | 71.43%  | 71.43%                   | 100%              |
| 90                | 100%    | 100%                     | 100%              |

Table 5 shows the highest accuracy occurring at random forest, which was in contrast with SVM and fuzzy kernel c-means that used in the paper of Rustam and Yaurita. This was recorded at a
level of 100%, when the percentage of training data is 70-90%, while others only achieved ~70% at 60-80%. If we average the result of each of combination model based on the percentage of training dataset, SVM gave us 71.93% of accuracy, fuzzy kernel c-mean was 71.35% and random forest was 98.4%.

4. Conclusion
Prediction of insolvency insurance has previously been conducted by Segovia-Vargas, et all using rough sets [1] and Rustam and Yaurita using support vector machines and fuzzy kernel c-means [2]. In this research presented a new approach to predict insurance insolvency by using random forest. The best model can achieve by tuning the hyperparameters, criterion (entropy and gini) and number of estimators (10, 50, 100) were used as a hyperparameters. The combination of hyperparameters gave us the best model while the criterion was entropy and number of estimators was 100, even random forest with entropy and gini shows similar results, although it only slightly performs better with entropy, that was 100% accuracy of training and 98.4% accuracy of testing in average. Subsequently, this technique was also able to predict with good accuracy, using 50% training data, which was in contrast with other methods [6]. This is very important, especially in the prediction of insolvency, where data is difficult to obtain. In comparison with past studies using the same data, random forest was observed to show better accuracy, at 100% for training, and also testing. For the next research, we suggest use another dataset that updated and bigger dimension.

Acknowledgment
This research was supported financially by Universitas Indonesia, with a DRPM UI PITTA 2018 research grant scheme.

References
[1] European Central Bank 2009 The Importance of Insurance Companies for Financial Stability Financial Stability Review
[2] P Ivan, C Dusan, and P Tatjana 2015 Role of Insurance Companies in Financial Market Faculty of Business Economics and Entrepreneurship, International Review No. 1-2.
[3] J M Ambrose and A M Carol 1994 Using Best’s Ratings in Life Insurer Insolvency Prediction The Journal of Risk and Insurance 61(2) 317-327
[4] R Bar-Niv and M L Smith 1987 Underwriting, Investment, and Solvency Journal of Insurance Regulation 5 409 – 428
[5] M J Segovia-Vargas, J A Gil Fana, A Heras-Martinez, J L Vilar-Zanon, and A Sanchis-Arellano 2002 Using Rough Sets to predict insolvency of Spanish non-life insurance companies
[6] Z Rustam and F Yaurita 2018 Insolvency Prediction Companies Using Support Vector Machines and Fuzzy Kernel C-Means
[7] R Pal 2017 Overview of predictive modeling based on genomic characterizations. Predictive Modelling of Drug Sensitivity. Predictive Modelling of Drug Sensitivity (pp. 121-148). https://doi.org/10.1016/B978-0-12-805274-7.00006-3
[8] Z Rustam and G S Saragh 2018 Predicting bank financial failures using random forest. In 2018 International Workshop on Big Data and Information Security, IWBIS 2018 (pp. 81-86). [8471718] (2018 International Workshop on Big Data and Information Security, IWBIS 2018). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/IWBIS.2018.8471718
[9] P R Patil and S A Kinarivala 2017 Automated Diagnosis of Heart Disease using Random Forest Algorithm International Journal of Advance Research, Ideas and Innovations in Technology 3(2) pp, 2452-132x
[10] Q Zou, K Qu, Y Luo, Y Ju, and H Tang 2018 Predicting Diabetes Mellitus with Machine Learning Techniques Frontiers in Genetics 9 pp. 515
[11] L Breiman 2001 Random forests Mach. Learn 45 pp. 5–32
[12] G Carvajal, M Maucec, and S Cullick 2018 Components of Artificial Intelligence and Data Analytics Intelligent Digital Oil and Gas Fields pp. 101-148
[13] W Mao and F Wang 2012 Cultural Modelling for Behavior Analysis and Prediction New Advances in Intelligence and Security Informatics.
[14] B Efron and R Tibshirani 1993 An Introduction to the Bootstrap (New York: Chapman & Hall)
[15] C Strobl 2008 Statistical Issues in Machine Learning – Towards Reliable Split Selection and Variable Importance Measures. Munchen.

[16] R Kohavi and F Provost 1998 On applied research in machine learning Columbia

[17] P Schratz, J Muenchow, E Iturritxa, J Richter, and A Brenning 2018 Performance evaluation and hyperparameter tuning of statistical and machine-learning models using spatial data Germany

[18] L Kaufman and P J Rousseeuw 1990 Finding Groups in Data, An Introduction to Cluster Analysis edited by John Wiley & Sons.

[19] V N Vapnik 1998 Statistical Learning Theory (New York Wiley)

[20] T Hastie, R Tibshirani, and J Friedman 2008 The element of statistical learning data mining, inference, and prediction (California: Springer)