INTI JOURNAL | eISSN:2600-7320
Vol.2020:06

Ensemble Learning Boosting Model of Improving Classification and Predicting

Bambang Siswoyo A[1], Nanna Suryana B[2], DA Dewi C[3]

[1] Computer Fakulty, Masoem University (MU), Indonesia
[2] Universiti Teknikal Malaysia Melaka (UTeM), Malaysia
[3] INTI International University, Malaysia

***Email:** bambangf1@gmail.com

Abstract

Artificial Intelligence Engineering is an important topic and has been studied extensively in various fields. Machine learning is part of Artificial Intelligence that has been used to solve prediction problems and financial decision making. An effective prediction model is one that can provide a higher prediction accurate, that is the goal of prediction model development. In the previous literature, various classification techniques have been developed and studied, which by combining several classifier approaches have shown performance over a single classifier. In building a boosting ensemble model, there are three critical issues that can affect model performance. First are the classification techniques actually used; the second is a combination method for combining several classifiers; and all three classifiers to be combined. This paper conducts a comprehensive study comparing the ensemble boosting classifier and three widely used classification techniques including AdaBoost, Gradient boosting, XGB Classifier. The results of the experiment with two financial ratio datasets show that the Ensemble Boosting Classifier has the best performance with an accurate value of 98%, while AdaBoost is 96%, Gradient_boosting is 98%, and XGB Classifier is 98%. Ensemble Boosting matches all available data, so the predict () function can be called to make predictions on new data.

Keywords

Ensemble Learning, Boosting, Financial Ratio, Classification

International Conference on Innovation and Technopreneurship 2020
Submission: 3 August 2020; Acceptance: 13 August 2020

Introduction

Ensemble Ensemble learning is a strategy where a group of models is used to solve problems that exist today, strategically ensemble learning combines various machine learning models into a single predictive model. In general, the ensemble method is mainly used to improve the accuracy of the overall performance of the model and combine some basic learners, for classification or prediction of the actual class. The more diverse the basic learner is, the stronger the final model will be. In each machine learning model, generalization error is given by the sum of the squares of bias + variance + irreducible error. By using the ensemble technique, it will reduce the bias and variance of the model. This reduces generalization errors overall.

Utilization of data in science applies to various fields to study hidden patterns and make predictions or descriptions accordingly, and it refers to the collection of techniques used to extract hidden knowledge, such as patterns, relationships, or rules from large data sets (Almasoud et al., 2015). This extracted knowledge can be analyzed and is able to predict future trends (Mhetre et al., 2017). Machine learning model optimization is an important step in producing more effective and efficient models. Optimization can include accelerating the data base reading process, determining parameters for the hypothesis function.

Machine learning applications are found in the retail, banking, military, health, financial, image, housing, etc. sectors, to achieve their goals, researchers develop different algorithms using expertise from various fields of study (Cherfi at al., 2018; Berquist et al. , 2017; Tsai et al., 2014; Hsu at al., 2012; Jardin at al., 2018; Hung at al., 2006; Sun et al, 2017; Brahmana et al., 2005; Hemmatfar, 2018; Zi et al. ., 2016; Zhao et al., 2017; Khairalla at al., 2018; Priya et al, .2018; Janggo, 2018; Santosh et al, 2020; Pisula et al, 2020; Zi et al, 2016; Barboza at al. , 2017; Altman, 2000; David, 2011; Chen, at all, 2013). This algorithm can be used to build models, which can obtain insights from previous data. This can be applied to solve problems related to classification, regression, grouping and optimization using algorithms such as decision trees, random forests, logistic regression, support vector machines (SVM), Naïve Bayes, K-Nearest Neighbor (KNN), K-Means and others..

In this paper, the proposed method is Ensemble Learning Gradient Boosting (EL-GB), where the EL-GB model has the advantage of being able to achieve better prediction and classification performance than other algorithms.

Methodology

The EL-GB model uses several base classifiers in the learning process, there are two stages in EL-GB learning. Stage 1, each base classifier used is trained using the same dataset so as to produce the results of their respective predictions. Stage 2, the meta classifier takes the prediction results from the base class as input to determine which class is most likely to test data.

Hyperparameter is the number of decision trees used in ensemble boosting. Decision trees are added to the model sequentially in an attempt to refine and improve the predictions made by

the previous trees. Thus, more trees is often better. The number of trees can be set via the n-estimators argument and the default is 100. The number of samples used for each tree can vary. This means that each tree fits into a randomly selected subset of the training data set. Using fewer samples introduces more variables for each tree, although it can improve the overall performance of the model. The number of samples used to match each tree is determined by the subsample argument and can be set to a fraction of the size of the training data set. The diagram of the ensemble boosting method used is shown in Figure 1. Ensemble Boosting Prediction.

This study uses a dataset of banking industry financial ratios that publish financial reports on the web site https://www.ojk.go.id as well as on the respective banking websites. The use of this dataset is intended for classification of financial performance based on the Altman Z-Score. The features used are as follows: capital to total assets; retained earnings to total assets; earning before interest and taxed to total assets; book value and book earning to total debt. While the output is Z-Score and target. The following is a summary of the financial ratio dataset used can be seen in Table 1. Financial Ratio Dataset. The method of measuring the results uses for measurement criteria, namely accuracy, precision, sensitivity and specificity of testing 10 fold cross validation.

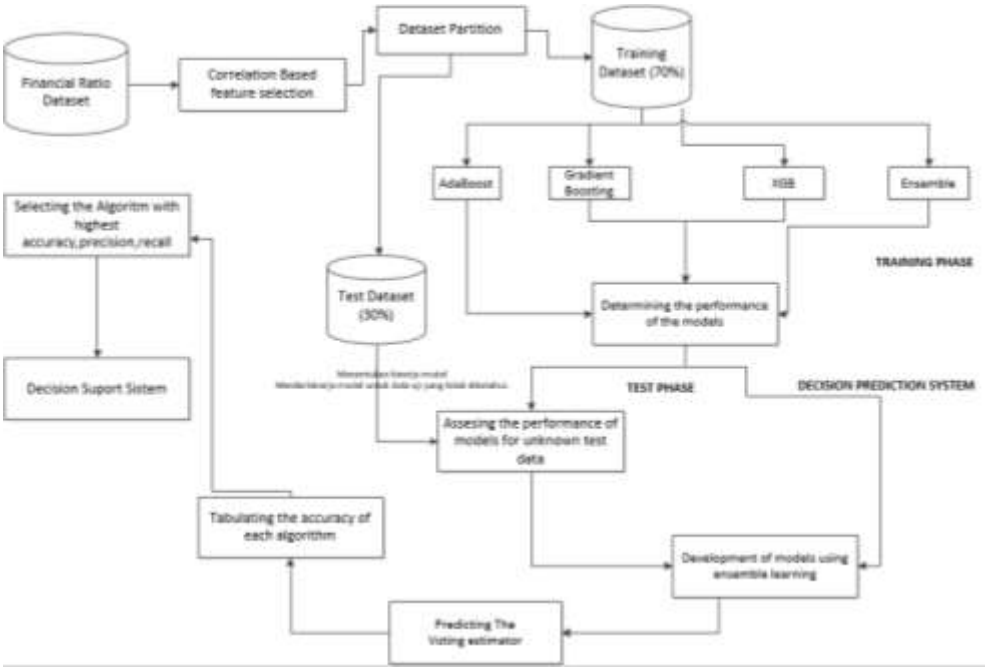

Figure 1. Boosting Clasifier Method

Table 1. Financial Ratio Dataset

WCTA	RETA	EBITA	BVEBDTA	ZSCORE	CLASS
4.2700	1.5000	0.6900	11.1000	0.4092	Distress Zone
.4376	0.1144	0.0610	0.2961	5.9081	Safe Zone
0.0024	0.0114	0.0177	0.0813	0.0852	Distress Zone
1.8000	1.5000	0.7200	31.4300	5.7425	Safe Zone
3.5900	1.6250	0.6100	27.6800	2.1127	Gray Zone

3.5200	1.8750	1.9000	13.1400	0.3796	Distress Zone
7.1000	1.9300	-1.8700	15.2700	1.2372	Distress Zone
1.8600	1.3000	1.3700	16.5400	0.0512	Distress Zone
6.8400	2.1200	0.1700	14.7600	1.1334	Gray Area
2.0900	1.7000	1.5000	11.5700	0.8147	Distress Area
0.1000	1.0000	0.8000	29.6000	1.2637	Gray Area
4.5700	2.0000	0.8800	11.9100	0.5932	Distress Zone
1.0200	1.3500	1.0300	20. 8300	1.7915	Gray Area
4.07000	2.000	0.27000	15.8500	0.6564	Distress Zone

The financial ratio dataset shows that there are four features as independent variables and two features as target classes. The graph is shown in Figure 2. Financial Ratio Dataset Graph

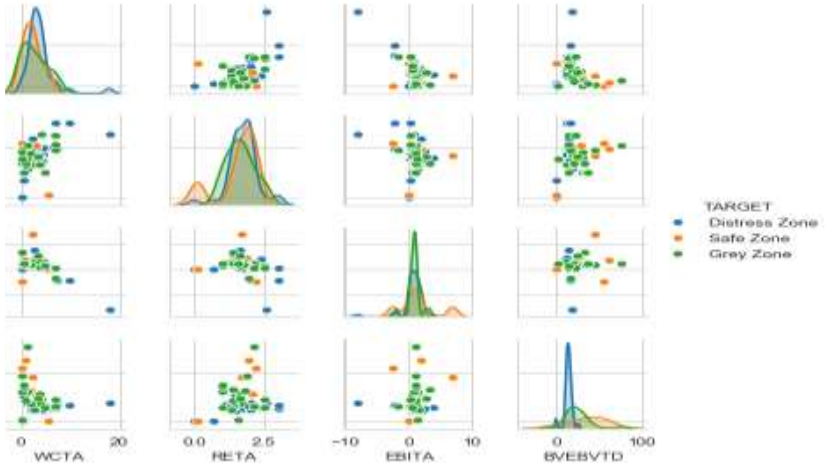

Figure 2. Financial Ratio Dataset Graph

Results and Discussion

The results of the AdaBoost Based Classifier training dataset with n_estimator 5, Gradient boosting with n-estimator 10, XGB with a max depth of 5, and a learning rate of 0.001, and ensemble using voting hard are shown in Table 2. Comparison Results of Distress Dataset Accuracy.

Table 2. Comparison Results of Distress Dataset Accuracy

Training Dataset	Ada Boost Clasifier %	Gradient Boost Classifier %	XG Boost Clasifier %	Ensemble Clasifier %
BankSaria	91	91	92	91
BankConven	96	98	98	98

Table 2 shows that the accuracy increases to 98% when using the boosting classifier based on the multiclass dataset, with majority voting, the Adaboost model is 96%, although with other

individual models the accuracy is just as good.

Table 3. Comparison Results of Confusion Matrix Accuracy

Training Dataset	Precision %	Recall %	F1-Score %
BankSaria	86	86	84
BankConven	86	86	84

Specific results may vary given the stochastic nature of the learning algorithm. In this case, we can see that the Boosting ensemble with hyperparameter the n estimator =10 achieves precision, recall, and F1-Score quite well on the four datasets as the ensemble boosting model can be used as the final model and make predictions for classification. First, ensemble Boosting matches all available data, then the predict () function can be called to make predictions on new data. A qualitative bankrupt dataset is a binary classification, the ensemble boosting model can evaluate on this dataset. The results of machine learning such as ensemble learning boosting show high potential for use in corporate bank distress finance prediction systems, especially when combined with knowledge of financial analysis. This paper strengthens research on the use of boosting that has been done by Santosh, (Santosh et al, 2020) with accuracy in ensemble boosting 98%. In this study, the level of accuracy with GradientBoost and XGBboost based learn shows the same accuracy with ensemble boosting, which is 98%, while AdaBoost is sensitive to noise data, this is greatly influenced by outliers because it tries to adjust each point perfectly so AdaBoost accuracy rate of 96%, lower compared to other based learn.

Conclusions

This study is designed to develop a learning ensemble boosting, to detect the level of control of the banking industry financial distress using a combination of based learning which refers to the performance of the model with the approach of using classification techniques; a combination method for combining several classifiers; and the number of classifiers to combine. The proposed model, compared with other classifiers, significantly improves accuracy, recall, precision, and F1Score. Ensemble boosting is easy to implement, iteratively corrects weak classifier errors and improves accuracy by combining weak based learns, can use many base classifiers, is not prone to overfitting. In future research, various objectives can be considered as follows: Building a server that functions to store banking financial reports so that it can integrate data, expert knowledge, feature selection, balance operations on the dataset, and add the use of various other influencing factors as a level of distress control financial factors such as liquidity factors and corporate governance indicator factors to improve the detection performance of the level of financial distress control by learning ensemble boosting.

Acknowledgment

We would like to thank Prof. Nanna Suryana Herman and Dr. Zuraida Binti Abal Abbas has supported in this study.

References

[1] A. M. Almasoud, H. S. Al-Khalifa, and A. Al-Salman, "Recent developments in data mining applications and techniques," in 2015 Tenth International Conference on Digital Information Management (ICDIM), 2015, pp. 36–42.

[2] Altman, E. I. (2000). Predicting Financial Distress of Companies : Revisiting The Z- Score and Zeta Models. Journal of Banking & Finance.

[3] A.Esteban,, R.N. Gracia, Matías, A. Elizondo, (2007) Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks, April 2008, Decision Support Systems 45(1):110-122,online

[4] Barboza. F, Herbert Kimurab , Edward Altmanc, (2017) Machine Learning Models and Bankruptcy Prediction, Expert Systems With Applications (2017). Doi: 10.1016/j.eswa.2017.04.006

[5] Bergquist SL, Brooks GA, Keating NL, Landrum MB, Rose S. Classifying lung cancer severity with ensemble machine learning in health care claims data. In: 2nd machine learning for healthcare conference. 2017. pp. 25–38.

[6] Brahmana, Rayenda K. 2005. Identifying Financial Distress Condition in Indonesia. *Birmingham Business School, University of Birmingham United Kingdom*

[7] C.Hsu and C. Lin, "Comparison of Methods for Multiclass Supporting Vector Machines," vol. 13, no. 2, p. 415-425, 2002.

[8] Cherfi, A., Nouira, K., and Ferchichi, A. (2018). Very Fast C4.5 Decision Tree Algorithm. Journal of Applied Artificial Intelligence, 2018,32(2), pp. 119-139

[8] Chih-Fong Tsai, Yu-Feng Hsu, Chih-Fong Tsai. A comparative study of classifier ensembles for bankruptcy prediction, 2014,Applied Soft Computing 24:977-984. DOI:10.1016/j.asoc.2014.08.047

[9] Chihli Hung, JingHongChen, Stefan Wermter. 2006. Hybrid Probability Based Ensembles For Bankruptcy Prediction.

[10]Diakomihalis, Mihail. 2012. The accuracy of Altman's models in predicting hotel bankruptcy. International Journal of Accounting and Financial Reporting 2: 96–113. [CrossRef]

[11]Du Jardin, Philippe. 2018. Failure pattern-based ensembles applied to bankruptcy forecasting. Decision Support Systems 107: 64–77. [CrossRef]

[12]Fedorova, Elena, Evgenii Gilenko, and Sergey Dovzhenko. 2013. Bankruptcy prediction for Russian companies:Application of combined classifiers. Expert Systems with Applications 40: 7285–93. [CrossRef]

[13]Khairalla MA, Ning X, AL-Jallad NT, El-Faroug MO. Short-term forecasting for energy consumption through stacking heterogeneous ensemble learning model. Energies. 2018;11:1–21. https://doi.org/10.3390/en11061605. Article Google Scholar

[14]Maknickiene N, Lapinskaite I, Maknickas A. Application of ensemble of recurrent neural networks for forecasting of stock market sentiments. Equilib Q J Econ Econ Policy. 2018;13:7–27. https://doi.org/10.24136/eq.2018.001. Article Google Scholar

[15]Mahmoud Hemmatfar, 2017. Prediction of firms' financial distress using adaboost algorithm and compairing its accuracy to artificial neural networks. Islamic Azad University.

[16]Myoung-Jong Kim, Dae-Ki Kang,2010. Ensemble with neural networks for bankruptcy prediction,Expert Systems with Applications 37(4):3373-3379. DOI: 10.1016/j.eswa.2009.10.012

[17] O. Purvinis, R., Virbickait˙e, P. ˇSukys. Klaip˙edos str, Panev˙eˇzys. 2008. Interpretable Nonlinear Model for Enterprise Bankruptcy Prediction. Nonlinear Analysis: Modelling and Control, 2008, Vol. 13, No. 1, 61–70

[18]Park, D., Yun, Y. and Yoon, M. (2012). Prediction of bankruptcy data using machine learning techniques. Journal of the Korean Data & Information Science Society, 23, 569-577. https://doi.org/10.7465/jkdi.2012.23.3.569

[19]Pisula. T, 2020. An Ensemble Classifier-Based Scoring Model for Predicting Bankruptcy of Polish Companies in the Podkarpackie Voivodeship, Department of Quantitative Methods, Faculty of Management, Rzeszow University of Technology, Poland.

[20]Priya P, Muthaiah U, Balamurugan M. Predicting yield of the crop using machine learning algorithm. Int J Eng Sci Res Technol. 2018;7:1–7.Google Scholar

[21]Platt, H., dan M. B. Platt. 2002. "Predicting Financial Distress". Journal of Financial Service Professionals, 56: 12-15.

[22]Santosh Shrivastava, P Mary Jeyanthi & Sarbjit Singh, (2020) Failure prediction of Indian Banks using SMOTE, Lasso regression, bagging and boosting, Cogent Economics & Finance, https://doi.org/10.1080/23322039.2020.1729569

[23]Sun, Jie, Hamido Fujita, Peng Chen, and Hui Li. 2017. Dynamic financial distress prediction with concept drift based on time weighting combined with AdaBoost support vector machine ensemble. Knowledge-Based Systems 120: 4–14. [CrossRef]

[24]Topaloglu, Zeynep. 2012. A Multi-period Logistic Model of Bankruptcies in the Manufacturing Industry. International Journal of Finance and Accounting 1: 28–37. [CrossRef]

[25]Tsai, Chih-Fong, Yu-Feng Hsu, and David C. Yen. 2014. A comparative study of classifier ensembles for bankruptcy prediction. Applied Soft Computing 24: 977–84. [CrossRef]

[26]Wen-Kuei Hsieh, Shang-Ming Liu, Sung-Yi Hsieh. Hybrid Neural Network Bankruptcy Prediction: An Integration of Financial Ratios, Intellectual Capital Ratios, MDA, and Neural Network Learning. Department of Finance, De Lin Institute of Technology, Taipei 236, Taiwan

[27]Zieba, Maciej, Sebastian K. Tomczak, and Jakub M. Tomczak. 2016. Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. Expert Systems with Applications 58: 93–101. [CrossRef]

[28]Zhao Y, Li J, Yu L. A deep learning ensemble approach for crude oil price forecasting. Energy Econ. 2017;66:9–16. https://doi.org/10.1016/j.eneco.2017.05.023. Article Google Scholar

[29]Zi ˛eba, Maciej, Sebastian K. Tomczak, and Jakub M. Tomczak. 2016. Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. Expert Systems with Applications 58: 93–101. [CrossRef]