Integration of Iterative Dichotomizer 3 and Boosted Decision Tree to Form Credit Scoring Profile

Alditama Agung Prasetyo, Budhi Kristianto
Fakultas Teknologi Informasi, Universitas Kristen Satya Wacana
budhik@uksw.edu

Abstract— Loan is becoming essential need in this modern life. Banks need to keep their NPL ratio low in order to maintain their financial health. One of customer's screening techniques is credit scoring. This studi is conducted to implement credit scoring profile using Integration of Iterative Dichotomizer 3 and Boosted Decision Tree. Decision tree is a simple method to classify a condition into two different classes using given classifier, and widely used to perform credit scoring in the financial industry. We integrated Iterative Dichotomizer 3 and Boosted Decision Tree methods and used Microsoft Azure Machine Learning tools to perform credit score profiling. This study is cross sectional in time and using 600 instances data of loan submission in Tangerang, Indonesia. The result shows good performance with performance evaluation metric of accuracy, precision, recall, and F1 score are 0.85, 0.885, 0.793 and 0.836 respectively.

Keywords— Boosted Decision Tree, Credit Scoring, Iterative Dichotomizer 3

I. INTRODUCTION

Loan is becoming essential need in this modern life. Almost in every financial needs, we may apply for loans. For example, car loan, home loan, business loan, and student loan. In some cases, debitors can't pay their loans back. Bank refers it as non-performing loan (NPL). The NPL ratio of Indonesia is one of the highest NPL ratio of ASEAN countries [1]. It was 2.73% as October 2019, compared to 2.2% of Philippines, Malaysia's 1.6%, 1.3% of Singapore, and 2% of Vietnam.

Banks need to keep their NPL ratio low in order to maintain their financial health. Screening and profile analysis for new customers are mandatory. One of screening techniques is credit scoring. Credit scoring is an efficient method to measure the systematic risk when financing the individual customers as well as the small and medium sized enterprises (SMEs) [2].

In this study, we will use iterative dichotomizer 3 and two-class boosted decision tree techniques to develop credit scoring method, and analyze its advantages compared to other decision tree techniques.

II. LITERATURE STUDIES

Basically, decision tree is a simple method to classify a condition into two different classes using given classifier. For example, we will classify balls into two classes called "big" and "small". We use classifier "if the diameter is under 10 cm, it called small. If the diameter is 10 cm or above, it called big". Figure 1 may help to figure out this understanding.

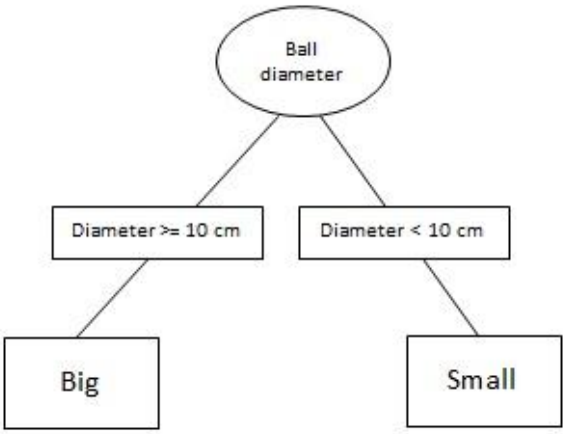

Figure 1. Decision tree method

Decision tree method was developed and expanded into many types to classify any specific conditions, including to develop credit score which used in lending and banking industries. A boosted decision trees method was used to develop credit scoring

ISSN 2442-7888 (online) DOI 10.24167/sisforma.v7i2.2659

model that help lenders decide whether to grant or reject credit to applicants [3]. Basically, boosted decision trees is a technique in which a result class from a decision tree is weighted to be developed as a new classifier to expand the tree and give more specific result based on more specific classifiers. Figure 2 may help to figure out this understanding.

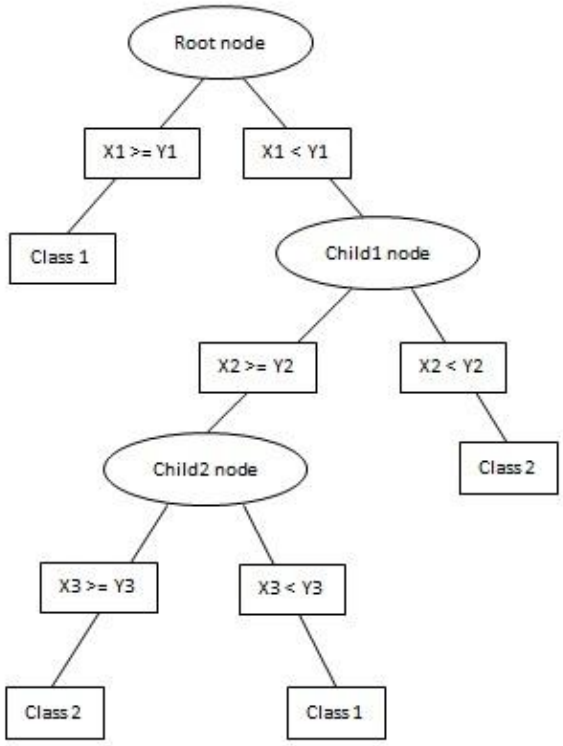

Figure 2. Boosted Decision Tree

Boosting is a procedure that aggregates many "weak" classifiers in order to build a new "strong" classifier. One of boosting techniques is AdaBoost or Adaptive Boosting proposed by Yoav Freund and Robert Schapire in 1996. The boosting process done by building a model from the training data, then creating a second model that attempts to correct the errors from the first model. This process repeated until the training data perfectly predicted. Each instance of the training dataset is weighted. The initial weight is set to "weight(xi) = 1/n" where xi is the i'th training instance and n is the number of training instances. According to Bastos, *boosted decision trees* outperformed the multilayer perceptron and the support vector machines on two real world credit card application datasets.

Another credit scoring analysis was conducted using integration between decision tree and neural network techniques called Decision Tree – Neuro Based Credit Risk Evaluation System [4]. They combined the advantages of decision tree such as easy to understood and fast learning, with the advantage of neural network such as capability to handle noised training data.

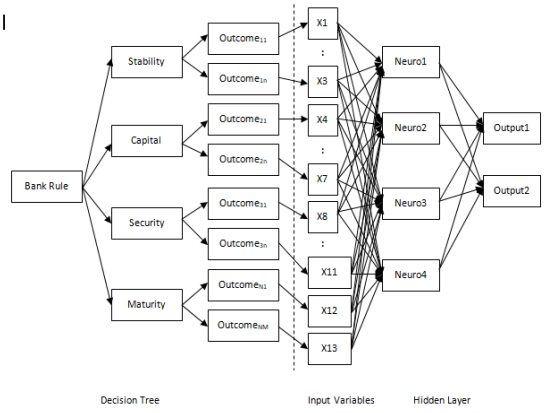

Figure 3. Credit scoring using decision tree – neuro based model

As we can see in figure 3, the decision tree technique was used to handles bank rules and criterions to give loan to customers, and the output was further processed with neural network technique to make final decision of the loan approval. They found that the accuracy rate of the decision tree – neuro based algorithm was 0.88, higher than decision tree's 0.68 and neural network's 0.75.

Iterative Dichotomizer 3 (ID3) Decision Tree also been used to develop credit scoring analyzer [5]. The Iterative Dichotomizer 3 (ID3) algorithm is used to create the shallowest decision trees possible and was invented by John Ross Quinlan in 1986. There are two different values that form the tree, entropy value and information gain value. Entropy value determines whether a node will be splitted (closer to 1) or not (closer to 0). When entropy value is zero, then it determines the class (leaf of tree). When entropy value closer to one, then the

ISSN 2442-7888 (online) DOI 10.24167/sisforma.v7i2.2659

attribute should be splitted and a new node will be formed by using the higher information gain value of the attributes.

Suppose we have dataset as seen in table 1. We can develop decision tree using ID3 algorithm as seen in figure 4.

Table 1. Dummy data for ID3 algorithm

Node	Attribute	Value	Approval	Entropy	Gain
1	Character	Good	Approved	0.564	0.732
		Bad	Not Approved	0.234	
	Address	Clear	Approved	0.875	**0.934**
		Unclear	Not Approved	0	
	Document	Complete	Approved	0.432	0.333
		Incomplete	Not Approved	0.354	
	Salary	Above 5K	Approved	0.123	0.543
		Bellow 5K	Not Approved	0.437	
2	Character	Good	Approved	0	**0.754**
		Bad	Not Approved	0	
	Document	Complete	Approved	0.644	0.387
		Incomplete	Not Approved	0.746	
	Salary	Above 5K	Approved	0.349	0.523
		Bellow 5K	Not Approved	0.531	

Another studies related to credit scoring also been conducted using neural network technique [6], segmentation technique [7], and fuzzy technique [8], [9], and [2].

III. METHODOLOGY

This study is cross sectional in time and using 600 instances data of loan submission in Tangerang, Indonesia. The data was normalized and had attributes CIFno, age, gender, region, income, marital status, number of child, car ownership, saving account ownership, checking account ownership, mortgage, and loan approval. The dataset then been processed by using Microsoft Azure Machine Learning with 480 instances data was used as training dataset. The accuracy and precision rate then be analyzed.

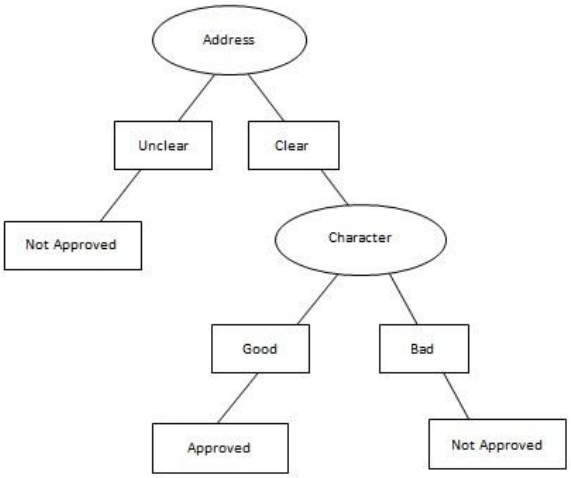

Figure 4. ID3 decision tree based on dummy data in table 1

IV. RESULTS AND DISCUSSION

Basically, Microsoft Azure named its boosted decision tree feature as two-class boosted decision tree. It only differentiates two-class boosted decision tree with multi-class boosted decision tree features, as two-class boosted decision tree is perfectly fits to binary classification problems and multi-class boosted decision tree may handle complex classification better.

First, we calculate entropy value for each instances data using formula

$$Entropy(S) = \sum_i^c - P_i \, Log2 \; P_i \quad \text{.....(1)}$$

Where:
C = number of attribute
Pi = number of instance

Then we calculate information gain value for each attributes to find the difference between entropy before split and average

ISSN 2442-7888 (online) DOI 10.24167/sisforma.v7i2.2659

entropy after split of the dataset based on given attribute values by using formula

$$Gain(A) = Entropy(S) - \sum_{i=1}^{c} \frac{|Si|}{S} Entropy(Si) ..(2)$$

Where:
C = number of attribute
A = particular attribute

For the attribute with many outcomes, information gain tends to be biased. That means it prefers the attribute with a large number of distinct values. Gain ratio handles the issue of bias by normalizing the information gain using Split Information. Split information can be calculated by using formula

$$Split\ Information\ A = -\sum_{i=1}^{j} \frac{Si}{S}\ log2\ \frac{Si}{S} \ldots..(3)$$

Where:
J = number of discrete values in attribute A
Si/S = the weight of the j-th partition

Then we calculate Gain Ratio by using formula

$$Gain\ Ratio\ (A) = \frac{Gain\ (A)}{Split\ Information\ (A)} \ldots\ldots\ldots..(4)$$

We performed Microsoft Azure Machine Learning calculation to form the trees using parameter as seen in table 2. After node is formed, we evaluate and boost the previous node to form next node by using AdaBoost algorithm.

Table 2. Parameter setup for Microsoft Azure two-class boosted decision tree

Parameter	Value
Create trainer mode	Single Parameter
Maximum number of leafs per tree	20
Minimum number of samples per leaf node	10
Learning rate	0.2
Number of trees constructed	100
Random number seed	blank

The Receiver Operating Characteristic (ROC) chart of the model is displayed in figure 5. A better model would have a higher True Positive Rate for the same False Positive Rate. As we can see in our ROC chart, we had around 0.87 of curve closer to the left, in which we had minimal false positive rate for our model.

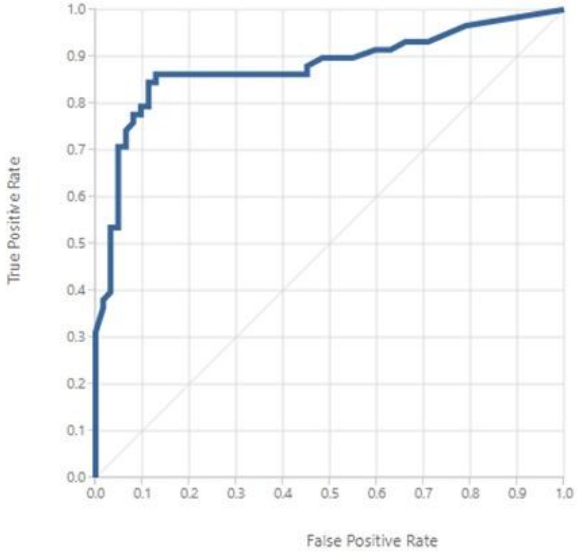

Figure 5. Receiver Operating Characteristic (ROC) chart of the model

V. MODEL PERFORMANCE EVALUATION

We used four metrics to evaluate the performance of our model, which are accuracy, precision, recall, and F1 score. Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. Accuracy can be calculated using formula

$$Accuracy = \frac{TP+TN}{(TP+FP+FN+TN)} \ldots\ldots\ldots\ldots\ldots..(5)$$

Where:
TP = True Positives
TN = True Negatives
FP = False Positives
FN = False Negatives

The second metric is precision, which is the ratio of correctly predicted positive observations to the total predicted positive observations. Precision can be calculated by using formula

$$Precision = \frac{True\ Positive}{(True\ Positive+False\ Positive)} \ldots\ldots..(6)$$

The third metric is recall (sensitivity), which is the ratio of correctly predicted positive observations to the all observations in actual class "YES". Good recall should

ISSN 2442-7888 (online) DOI 10.24167/sisforma.v7i2.2659

have value of 0.5 and above. Recall can be calculated as

$$Recall = \frac{True\ Positive}{(True\ Positive + False\ Negative)} \quad \ldots\ldots\ldots(7)$$

And the forth metric is F1 Score, which is the weighted average of precision and recall. Therefore, this score takes both false positives and false negatives into account. If we have an uneven class distribution, F1 score gives us better look rather than accuracy, while accuracy works best if false positives and false negatives have similar cost. F1 score can be calculated as

$$\mathrm{F1} = 2 \times \frac{Precision * Recall}{Precision + Recall} \quad \ldots\ldots\ldots\ldots\ldots\ldots(8)$$

Figure 6 shows complete performance of our model.

True Positive	False Negative	Accuracy	Precision
46	12	0.850	0.885
False Positive	**True Negative**	**Recall**	**F1 Score**
6	56	0.793	0.836
Positive Label	**Negative Label**		
YES	NO		

Figure 6. Model performance evaluation

The result shows that our mixed method yields high accuracy, precision, recall, and F1 score of 0.85, 0.885, 0.793 and 0.836 respectively.

Compared to other method like neural network, decision tree is still better fits to handle binary classification such as credit scoring, since it only has two class for the final output, approved and rejected. Although neural network have capability to handle complex attributes and scenarios, it still has limitation, especially due to its black box nature, which is difficult to explained [6].

Combination of decision tree and neural network may perform better performance, especially for cases with multi-class attributes [4]. The decision tree part may provide clear and distinct perceptrons for neural network part. It made the model much more adaptive to handle complex decision making cases.

VI. LIMITATION AND FUTURE WORK

We integrated Iterative Dichotomizer 3 and Boosted Decision Tree methods to form credit scoring profile and the result shows good performance on this technique. However, the boosted decision tree is one of the memory-intensive learners and the current implementation uses relatively high amount of memory. Therefore, we suggest to continue this research and combine with another method to minimize this limitation and to gain better performance of the larger dataset handling.

REFERENCES

[1] Kontan. (2019), "NPL Indonesia salah satu tertinggi di Asean, begini kata pengamat (Indonesia's NPL is one of the highest in Asean, according to observer)", available at https://keuangan.kontan.co.id/news/npl-indonesia-salah-satu-tertinggi-di-asean-begini-kata-pengamat (accessed 1 May 2020).

[2] Ilter, D. and Kocadagli, O. (2019), "Credit scoring by artificial neural networks based cross-entropy and fuzzy relations". Sigma Journal of Engineering and Natural Sciences, Vol. 37 No. 3, pp.855-870.

[3] Bastos, J. (2008), "Credit scoring with boosted decision trees", working paper, Munich Personal RePEc Archive, Munich, Germany.

[4] Kabari, L. and Nwachukwu, E. (2013). "Credit risk evaluating system using decision tree – neuro based model". International Journal of Engineering Research & Technology, Vol. 2 No. 6, pp.2738-2745.

[5] Ilayani, Nangi, J. and Pasrun, Y. (2018). "Aplikasi data mining untuk penilaian kredit menggunakan decision tree algoritma ID3 studi kasus PT. Mandala Multi Finance cabang Kendari (Data mining application for credit scoring using ID3 decision tree, case study of Mandala Multi Finance

ISSN 2442-7888 (online) DOI 10.24167/sisforma.v7i2.2659

Corporation Kendari branch)". semanTIK, Vol. 4 No. 1, pp.65-76.

[6] Baesens, B., Setiono, R., Mues, C., and Vanthienen, J. (2003). "Using neural network rule extraction and decision tables for credit-risk evaluation". Management Science, Vol. 49 No. 3, pp.312-329.

[7] Bijak, K. and Thomas, L. (2012). "Does segmentation always improve model performance in credit scoring?". Expert Systems with Applications, Vol. 39, pp.2433-2442.

[8] Tsipouras, M., Exarchos, T., and Fotiadis, D. (2008). "A methodology for automated fuzzy model generation". Fuzzy Sets and Systems, Vol. 159, pp.3201–3220.

[9] Wahyuningtyas, G., Mukhlash, I., and Soetrisno. (2014). "Aplikasi data mining untuk penilaian kredit menggunakan metode Fuzzy Decision Tree (Data mining application for credit scoring using Fuzzy Decision Tree method)". Jurnal Sains dan Seni POMITS, Vol. 2 No.1, pp.1-6.

[10] Abdou, H. and Pointon, J. (2011). "Credit scoring, statistical techniques and evaluation criteria: A review of the literature". Intelligent Systems in Accounting, Finance & Management, Vol. 18 (2-3), pp.59-88.

[11] Eddy, Y. and Bakar, E. (2017). "Credit scoring models: Techniques and issues". Journal of Advanced Research in Business and Management Studies, Vol. 7, No. 2, pp.29-41.

[12] Faddoul, J., Chidlovskii, B., Gilleron, R., and Torre, F. (2012). "Learning multiple tasks with boosted decision trees". Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Bristol, United Kingdom, pp.681-696.

[13] Madadipouya, K. (2015). "A new decision tree method for data mining in medicine". Advanced Computational Intelligence: An International Journal (ACII), Vol.2 No.3, pp.31-37.

[14] Melissa, I. and Oetama, R. (2013). "Analisis data pembayaran kredit nasabah bank menggunakan metode data mining (Analysis of bank's customer credit payment using data mining method)". ULTIMA InfoSys, Vol. 4 No. 1, pp.18-27.

[15] Ponomareva, N., Colthrust, T., Hendry, G., Haykal, S., and Radpour, S. (2017). "Compact multi-class boosted trees". 2017 IEEE International Conference on Big Data, Boston, MA, USA.

[16] Rahman, S., Irfan, M., Raza, M., Ghori, K., Yaqoob, S., and Awais, M. (2020). "Performance analysis of boosting classifiers in recognizing activities of daily living". International Journal of Environmental Research and Public Health, Vol. 17 No. 3, pp.1082-1097.

ISSN 2442-7888 (online) DOI 10.24167/sisforma.v7i2.2659