Ensemble GradientBoost for Increasing Classification Accuracy of Credit Scoring

Armin Lawi Department of Computer Science Universitas Hasanuddin Makassar, Indonesia armin@unhas.ac.id

Firman Aziz*) Post-graduate Electrical Engineering Universitas Hasanuddin Makassar, Indonesia firmanaziz88@gmail.com

Syafruddin Syarif Department Electrical Engineering Universitas Hasanuddin Makassar, Indonesia ssyariftuh@gmail.com

***Abstract*— The method for Credit Scoring has been developed to select a better model in predicting credit risk. Data mining methods are superior to the statistical methods of dealing with Credit Scoring issues, especially for nonlinear relationships between variables. By flashing the ensemble method with statistical methods, proven to achieve a higher level of accuracy than the method of data mining. This paper proposes a credit scoring algorithm using Ensemble Logistic Regression by boosting the method using the GradientBoost algorithm. Two datasets for implementing the algorithm, i.e., German and Australian Dataset. The results showed that GradientBoost Ensemble managed to improve the performance of a single classification Logistic Regression and achieve the highest level of accuracy in both datasets. The proposed method produces accuracy of 81% for German datasets and 88.4% for Australian datasets.**

Keywords—Credit Scoring; Logistic Regression; Ensemble Bagging; Ensemble GradientBoost.

I. INTRODUCTION

The current development of the credit industry is very rapid. Lending is a profitable business activity for banks and other financial institutions. The high demand for credit in a bank does not make the bank will be able to grant all existing applications. Therefore, banks need to conduct a selection process to see which borrowers are eligible to be loaned. The process is an appraisal process using historical data from prospective borrowers to then be classified to make decisions worthy of a loan or not. A prospective debtor who passes the selection is called good debtor, and his credit request will be granted. On the contrary, prospective borrowers who do not pass the selection are called bad debtors and their credit applications will be rejected. The process is called credit scoring.

Credit Scoring is described as a statistical technique for converting data in making credit decisions [1]. According to [2]., Credit Scoring is a method of measuring the inherent risk to customers by analyzing customer data to determine the likelihood that prospective borrowers will default on the loan.

The most popular classification method adopted in the credit scoring industry is logistic regression that is relatively easy to understand and implement [3], [4], [5]. However, some studies use data mining methods such as Ant Colony Optimization [6], Genetic Programming [6], [7], Support Vector Machine [4], [8], [9], [10], [11], [12] and Artificial neural network [4], [13], [14], [15]. It was concluded that the method of data mining is superior to the statistical method in handling Credit Scoring problem, especially for nonlinear relationships between variables and generalizability skills in building models. Although statistical methods are only suitable for situations where the relationship of the underlying variable is linear when in the ensemble, it achieves the highest degree of accuracy compared to the method of data mining [16].

The ensemble method is based on some training to solve the same problem and then the output of a single classification combined with the ensemble method into one classifier to provide improved performance [17]. Research [18] performs ensemble bagging and boosting by using ANN as a single classification. ANN ensemble shows accurate results and low generalization errors. In [19] it examines the use of SVM and KNN as a basic classifier and uses ensemble bagging and boosting to improve the performance of the basic classification. The results show the advantage of the ensemble in terms of classification performance. In [20] developed an ensemble model based on LS-SVM to reduce bias and improve classification accuracy. Research [16] proposes to develop and introduce a systematic credit rating model based on homogeneous and heterogeneous classifier ensembles based on three LR, ANN and SVM classifiers. The results show that the heterogeneous ensemble classifier provides better predictive performance than homogeneous and singular clustering in terms of accuracy.

In this study proposes Credit Scoring identification using Ensemble Logistic Regression with Boosting method. Boosting usage is preferred because it focuses on misclassified issues and has a tendency of increased accuracy compared to Bagging. Ensemble GradientBoost classification results will be compared with a single logistic regression and ensemble Bagging classification to see the accuracy of each method.

*) Corresponding Author: firmanaziz88@gmail.com

II. PRELIMINARIES

A. Credit Scoring

Credit Scoring is a classification problem that classifies potential borrowers into two classes of 'good' borrowers and 'bad' borrowers based on the characteristics of prospective borrowers [21], e.g., age, economic conditions, social status, guarantees, etc. The purpose of the credit evaluation process is to reduce the risk of 'bad' customers for credit. The result of classification is an important process in the corporate credit management toolkit. Thus, Credit Scoring is an important technology for banks and other financial institutions as it seeks to minimize risk.

B. Classification using Logistic Regression Techniques

Logistic regression is statistic method widely used to solve classification problem and regression [22]. Logistic Regression is used to make example of binary result variable, usually, it is represented by 1 or 0. The scoring model must binary (accept / good loan, 1 or reject / bad credit, 0) and this depends on the category of the independent variable [23]. The function of the classifier shown by equation:

$$log\left(\frac{p}{1-p}\right) = \sum_{i=1}^{n} \beta^{(i)} * x^{(i)} + \varepsilon = \beta^T x + \varepsilon \qquad (1)$$

Where $\beta=(\beta^{(1)}, \beta^{(2)},...,\beta^{(n)})$ is coefficient vector of hyperplane. The probability of customer stop on equation (1) can be simply formulated as in the following eqution (2).

$$p = \frac{e^{\beta^T x+\varepsilon}}{1+e^{\beta^T x+\varepsilon}} \qquad (2)$$

C. Classification Ensembles

The ensemble method can reduce classification errors effectively, and is believed to perform well compared to the use of a single classifier. The main idea of the ensemble method is to combine several sets of models that solve a similar problem to obtain a more accurate model [20]. Compared to an individual classifier, they only learn and train a set of data only. But ensemble classifiers learn and train the various data generated from the original dataset and the results will build a set of hypotheses from the data trained and produce better accuracy [24].

Some ensemble classifiers techniques have been developed such as bagging, boosting, random forest and rotation forest. Due to space limitations, only the bagging and boosting methods will be explained.

- *Bagging*

Bagging is one of the first ensemble learning algorithms. Easy to implement and deliver good performance results. Diversification of data to be trained achieved by making the number of different bags (n-bag). Each bag is filled with randomly generated data from the training dataset[16].

Bagging (Bootstrap Aggregating) algorithm creates M bootstrap samples $T_1, T_2, \ldots, T_M$ randomly drawn from the original training set T, of size n[23]. Each bootstrap sample T_t of size n is then used to train a base classifier C_t. Predictions on new observations are made by taking the majority vote of the ensemble C^* built from $C_1, C_2, \ldots, C_M$[25].

Algorithm for ensemble Bagging
Given training set of size n and base classification algorithm $C_t(x)$.

1. Input sequence of training samples $(x_1:y_1), \ldots (x_n:y_n)$ with labels $y \in Y = (-1,1)$
2. Initialize probability for each example in learning set $D_1(i) = \frac{1}{n}$ and set $t = 1$.
3. Loop while $t < B = 100$ ensemble members
   a. Form training set of size n by sampling with replacement from distribution D_t
   b. Get hypothesis $ht: X \rightarrow Y$
   c. Set $t = t + 1$

   End of loop
4. Output the final ensemble hypothesis

$$C^*(x_i) = h_{final}(x_1) = argmax \sum_{t=1}^{B} I(C_t(x) = y).$$

- *Boosting*

Boosting is a common and effective method for building an accurate classifier by combining weak classifiers. The use of boosting is preferred because it focuses on misclassified issues and has a tendency of increased accuracy compared to the bagging method.

The focus of this method is to generate a series of base classifiers. The training sets used for each base classifier are selected based on the performance of the previous classifiers. In boosting, samples that are not predicted correctly by the classifiers will be selected then the predicted samples correctly. Boosting tries to produce a new base classifiers that are better for predicting samples that in previous base classifiers have poor performance [24]. One of the most popular algorithms of the boosting method is the GradientBoost algorithm.

- *GradientBoost*

Gradient boosting is a machine learning technique for regression and classification problems by matching simple parameter functions that produce predictive models in the form of ensembles of weak models [26].

Algorithm Gradient Boosting

1. $F_0(x) = \arg\min_\gamma \sum_{i=1}^{N} \Psi(y_i, \gamma)$.
2. for $m = 1$ to M do:
3. $\quad \tilde{y}_{im} = -\left[\frac{\partial \Psi(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}, i = 1, N$
4. $\quad \{R_{lm}\}_1^L = L - \text{terminal node tree}(\{\tilde{y}_{im}, x_i\}_1^N)$
5. $\quad \gamma_{lm} = \arg\min_\gamma \sum_{x_i \in R_{lm}} \Psi(y_i, F_{m-1}(x_i) + \gamma)$
6. $\quad F_m(x) = F_{m-1}(x) + \nu \cdot \gamma_{lm} 1(x \in R_{lm})$
7. endfor

In the case of estimation functions one system consists of a random 'output' or 'response' variable y and a set of random 'input' variables $x = \{x_1, \ldots, x_n\}$. Given the 'training' of samples $\{y_i, \mathbf{X}_i\}_1^N$ of known values (y; x), the goal is to find

the function $F*(x)$ that maps x to y, so through distribution to all values (y; x) [26].

D. Performance Evaluation

Performance evaluation of classification method can be seen from the level of classification error. To count mark of classification error can use confusion matrix, the confusion matrix is usually called with contingency table like on Table I.

Table I. *Confusion Matrix/ Contingency Table*

Actual/Prediction	Good Loans	Bad Loans
Good Loans	TP	FN
Bad Loans	FP	TN

If TN, FP, FN, and TP are evaluated, then the accuracy level of the model can be given with the following equation (3).

$$accuracy = \frac{TP+TN}{TN+FP+FN+TP} \quad (3)$$

III. Experimental

A. Dataset

In this research, the data used to evaluate the accuracy of the proposed model are the German dataset and the Australian dataset divided into 70% for training data and 30% for data testing. This dataset has been widely used by researchers in credit scoring literature and is available in UCI machine learning repository. Generally, the dataset can be seen in Table II.

Table II. *Description of dataset*

Dataset	Loans	Good Loans	Bad Loans	Attributes
German	1000	700	300	20
Australian	690	307	383	14

B. Normalization Data

Each attribute in the dataset has different values for eliminating data redundancy and building the data model in the same interval. The data must be transformed from a scale of values different from the general scale. The dataset attribute is normalized with values between 0 and 1. This transformation is done by taking the maximum value of each attribute and dividing all values in the attribute with the maximum value.

C. Implementation

The focus of this research is to propose an ensemble Logistic Regression using the Boosting method with Gradient Boosting algorithm to improve the performance of a single Logistic Regression classification. Systematically the proposed model has the following stages:

Algorithm Ensemble Logistic Regression Using GradientBoost.

- Load dataset
- Identify label attribute and class
- Determine training set and testing set
- Form classification
- Initial model
  In this research used classification of logistic regression with maximum likelihood approach
- Use function and likelihood logarithm
- Differentiate likelihood similarity
- Literacy m=1 to M
- Count pseudo-residuals
- Train pseudo-residuals use training set
- Count multiplier
- Renew model
- Output model

IV. Result

All results are based on the testing process using the German dataset and Australian dataset.

Table III. *Classification results*

Classification	Ensemble	Accuracy German Dataset	Accuracy Australian Dataset
Logistic Regression	-	77 %	85.9 %
	Bagging	78.3 %	86.4 %
	GradientBoost	81 %	88.4 %

Table III. shows the classification results on the German dataset and Australian dataset. The single logistic regression classification model produces 77% accuracy on the German dataset and 85.9% on the Australian dataset. Level of accuracy based on an estimation of Maximum Likelihood Estimation (MLE) parameter of Logistic Regression.

Ensemble Logistic Regression using Bagging method yields 78.3% accuracy on German dataset and 86.4% on Australian dataset.

Next shows the results of the proposed model that is Ensemble Logistic Regression with boosting method using a GradientBoost algorithm.

The ensemble GradientBoost achieves a higher degree of accuracy in both datasets. For the German dataset yielded 81% accuracy and 88.4% accuracy on Australian datasets. Level of accuracy based on an estimation of Maximum Likelihood Estimation (MLE) parameter usage, learning rate will be set = 0.5 and the maximal value of single classification (max depth) will be set = 4 to get the best value depending on the interaction of the input variable.

V. Conclusion

In this research, proposed Ensemble Logistic Regression model with Boosting method using GradientBoost algorithm to see a performance in terms of accuracy level. The Ensemble GradientBoost successfully improves the performance of a single Logistic Regression classification and achieves a higher degree of accuracy across both datasets. For the German dataset yields 81% accuracy and 88.4% of Australian dataset.

REFERENCES

[1] D. J. Hand, S. Young, and Y. Kim, "Optimal bipartite scorecards," vol. 29, pp. 684–690, 2005.

[2] H. A. Abdou, "Credit Scoring , Statistical Techniques and Evaluation Criteria : A Review of the Literature," vol. 18, pp. 59–88, 2011.

[3] J. Vijay S. Desai, Jonathan N. Crook, George A. Overstreet, "A comparison of neural networks and linear scoring models in the credit union environment," vol. 2217, no. 95, 1996.

[4] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen, "Benchmarking state-of-the-art classification algorithms for credit scoring," pp. 627–635, 2003.

[5] T. Lee and I. Chen, "A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines," vol. 28, pp. 743–752, 2005.

[6] R. Aliehyaei and S. Khan, "Ant Colony Optimization , Genetic Programming and a Hybrid Approach for Credit Scoring : A Comparative Study," 2014.

[7] K. Tran and T. Duong, "Credit Scoring Model : A Combination of Genetic Programming and Deep Learning," no. December, pp. 145–149, 2016.

[8] C. Huang, M. Chen, and C. Wang, "Credit scoring with a data mining approach based on support vector machines," vol. 33, pp. 847–856, 2007.

[9] Y. Wang, S. Wang, and K. K. Lai, "Evaluate Credit Risk," vol. 13, no. 6, pp. 820–831, 2005.

[10] T. Bellotti and J. Crook, "Support vector machines for credit scoring and discovery of significant features," Expert Syst. Appl., vol. 36, no. 2, pp. 3302–3308, 2009.

[11] T. Harris, "Expert Systems with Applications Credit scoring using the clustered support vector machine," Expert Syst. Appl., vol. 42, no. 2, pp. 741–750, 2015.

[12] B. Yi and J. Zhu, "Credit Scoring with an Improved Fuzzy Support Vector Machine Based on Grey Incidence Analysis," pp. 173–178, 2015.

[13] D. West, "Neural network credit scoring models," vol. 27, 2000.

[14] A. F. Atiya and S. Member, "Bankruptcy Prediction for Credit Risk Using Neural Networks : A Survey and New Results," vol. 12, no. 4, pp. 929–935, 2001.

[15] A. Khashman, "Expert Systems with Applications Neural networks for credit risk evaluation : Investigation of different neural models and learning schemes," Expert Syst. Appl., vol. 37, no. 9, pp. 6233–6239, 2010.

[16] M. Ala, "A systematic credit scoring model based on heterogeneous classifier ensembles," 2015.

[17] C. Tsai, "Combining cluster analysis with classifier ensembles to predict financial distress," Inf. Fusion, vol. 16, pp. 46–58, 2014.

[18] D. West, S. Dellana, and J. Qian, "Neural network ensemble strategies for ÿnancial decision applications," vol. 32, pp. 2543–2559, 2005.

[19] L. Nanni and A. Lumini, "An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring," Expert Syst. Appl., vol. 36, no. 2, pp. 3028–3033, 2009.

[20] L. Zhou, K. Keung, and L. Yu, "Expert Systems with Applications Least squares support vector machines ensemble models for credit scoring," Expert Syst. Appl., vol. 37, no. 1, pp. 127–133, 2010.

[21] Y. S. Kim and S. Y. Sohn, "Managing loan customers using misclassification patterns of credit scoring model," vol. 26, pp. 567–573, 2004.

[22] S. Akkoç, "An empirical comparison of conventional techniques , neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis : The case of Turkish credit card data," vol. 222, pp. 168–178, 2012.

[23] L. C. Thomas, "A survey of credit and behavioural scoring : forecasting financial risk of lending to consumers," vol. 16, pp. 149–172, 2000.

[24] Ludmila I. Kuncheva, Combining Pattern Classifiers Methods and Algorithms. 2004.

[25] A. I. Marqués, V. García, and J. S. Sánchez, "Expert Systems with Applications Exploring the behaviour of base classifiers in credit scoring ensembles," vol. 39, pp. 10244–10250, 2012.

[26] B. J. H. Friedman, "1999 REITZ LECTURE," vol. 29, no. 5, pp. 1189–1232, 2001.