Creditor Classification Logistic Regression Ensemble Boosting and Logistic Regression in Creditor Classification with Binary Response

ABELA CHAIRUNISSA, SOLIMUN SOLIMUN, ADJI ACHMAD RINALDO FERNANDES
Department of Statistics, Faculty of Mathematics and Science
Brawijaya University
Jl. Veteran, Malang 65145 East Java
INDONESIA

Abstract: - Credit risk is the risk that has the greatest opportunity to occur in banking. The number of bad loans will also affect bank performance. The banking sector needs to know whether a prospective creditor is classified as a risky person or not. The purpose of this study is to classify creditors and compare the classification results through logistic regression with the maximum likelihood model and the Boosting algorithm, especially the AdaBoost algorithm, and to select a model with the Boosting algorithm Credit Scoring aims to classify prospective creditor into two classes, namely good prospective creditor (Performing Loan) and bad prospective creditor (Non Performing Loan) based on certain characteristics. The method often used for classifying creditor is logistic regression, but this method is less robust and less accurate than data mining. Thus, there is a need for methods that provide greater accuracy. Among the methods that have been proposed is a method called Boosting, which operates sequentially by applying a classification algorithm to the reweighted version of the training data set. This study uses 5 datasets. The first dataset is secondary data originating from data on non-subsidized homeownership creditors of Bank X Malang City. While the other datasets are simulation data with many samples of 10, 500, and 1000. The results of this study indicate that ensemble boosting logistic regression is more suitable for describing binary response problems, especially creditor classification because it provides more accurate information. For high-dimensional data, which is represented by a sample size of 10, ensemble logistic regression is proven to be able to produce fairly accurate predictions with an accuracy rate of up to 80%, whereas in the logistic regression analysis the model raises N.A because many samples < many independent variables. The use of boosting is preferred because it focuses on problems that are misclassified and have a tendency to increase to higher accuracy.

Key-Words: - AdaBoost, Boosting, Credit, Credit Scoring, Ensemble Logistic Regression, Logistic Regression

Received: June 22, 2021. Revised: November 27, 2021. Accepted: December 10, 2021. Published: December 21, 2021.

1 Introduction

Classification is a statistical method that can be used to classify data arranged systematically. Lots of classification methods have been found, one of which is logistic regression. Logistic regression describes the relationship between dependent and independent variable which has two or more categories (Hosmer and Lemeshow, 1989). Logistic regression is a statistical method to describes the relationship between dependent and independent variable which has two or more categories (Hosmer and Lemeshow, 1989).

In this study, logistic regression will be applied to credit scoring. Credit Scoring aims to classify prospective creditor into two classes, namely good prospective creditor (Performing Loan) and bad prospective creditor (Non Performing Loan) based on certain characteristics. Traditional methods, such as logistic regression, are usually used in these situations, but they are less robust and accurate. This method does not work very well when there are interruptions in the data. As the complexity of these problems increases, there is a need for methods that provide greater accuracy. One of them is data mining. Data mining helps data analysis to be faster, more accurate, and cheaper.

Creditor classification data is unbalanced data. According to Weiss (2013), the class imbalance is the presence of an unbalanced number between classes contained in a dataset. There have been many studies that have developed credit scoring to classify creditors and potential creditors. Among the methods that have been proposed is a method called Boosting, which operates sequentially by applying a classification algorithm to the reweighted version of the training data set. Boosting is a form of the ensemble in logistic regression.

The advantages of a logistic regression ensemble, when compared to logistic regression are

when the data used are of high dimensions or if the number of predictor variables is more than the number of samples. For high-dimensional data, logistic regression will produce inaccurate predictions because several problem arise, so that the logistic regression ensemble is a solution for classifying high-dimensional data. The use of boosting is preferred because it focuses on misclassified problems and has a tendency to increase in higher accuracy.

Credit risk assessment is an important thing to do for banks. The quality of provision of funds and readiness to face loss risk greatly affects the performance and sustainability of rural banks (BI, 2006). If creditors in a credit bank experience default, the credit bank will suffer losses and reduce the capital they have. The existence of a credit risk assessment aims to anticipate the default. Although some credit risks cannot be avoided, banks can anticipate them in several ways. For example, creditors who have a high-risk level must have a higher income and be given a higher interest rate than creditors with a lower risk level. Furthermore, the granting of credit decisions must be guaranteed, will the creditor provide high returns or be too risky to be given credit.

Based on these problems, this study aims to classify creditors using Ensemble Logistic Regression with the Boosting method, namely AdaBoost. Then, the classification results of Ensemble Logistic Regression and Classical Logistic Regression will be compared to see the level of accuracy of each method.

2 Literature Review

2.1 Classification

Classification is a process to get a model or function that can distinguish classes in data. According to Johnson and Wichern (2007), the classification procedure is an evaluation to see the possibility of misclassification by a classification function. A good classification procedure is determined by a small misclassification value. One important thing to produce a classification procedure is to calculate the error rate or probability of misclassification (misclassification). There is a measuring tool that can be used to determine misclassification that does not depend on the distribution of the population and can simplify the calculation of various classification procedures.

Classification is the process of finding a model or function that explains or distinguishes a concept or data class to estimate the unknown class of an object. In classifying data there are two processes carried out, namely:

1. Training

   In the training process, a training set with known labels is used to build a model or function.

2. Testing

   To determine the accuracy of the model or function that will be built in the training process, data called a testing set is used to predict the labels.

2.2 Logistic Regression

According to Hosmer and Lemeshow (2000), the purpose of analyzing categorical data using logistic regression is to get the best and simplest model to explain the relationship between the outputs of the response variables (Y) with its predictor variables (X). Response variables in logistic regression can be categorical or qualitative, while predictor variables can be qualitative and quantitative. If the variable Y is a binary variable or dichotomy in the sense that the response variable consists of two categories, namely "success" (Y = 1) or "fail" (Y = 0), then the variable Y follows the Bernoulli distribution which has the probability density function:

$$f(y_i) = \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i} \quad , \quad y_i = 0,1 \tag{1}$$

So that it is obtained:

For $y_i = 0$ then

$$f(0) = \pi(x_i)^0(1-\pi(x_i))^{1-0} = 1-\pi(x_i)$$

For $y_i = 1$ then

$$f(1) = \pi(x_i)^1(1-\pi(x_i))^{1-1} = \pi(x_i)$$

Suppose the probability of the response variable for a given value x, denoted as $\pi(x)$. The general model $\pi(x)$ is denoted as follows:

$$\pi(x) = \frac{\exp(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p)}{1-\exp(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p)}, \tag{2}$$

then the estimator $\beta = (\beta_0, \beta_1, \ldots, \beta_n)$ using the maximum likelihood method equation is the solution of the likelihood equation:

$$\sum_{t=1}^{n}(Y_t - \pi(x_i)) = 0 \tag{3}$$

and

$$\sum_{j=1}^{p}\sum_{i=1}^{n} x_{ij}\left(Y_i - \pi(x_i)\right) = 0 \tag{4}$$

The form $\frac{\partial \ell(\beta)}{\partial \beta}$ in the previous equation can be written as:

$$\frac{\partial \boldsymbol{\ell}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \boldsymbol{x}(\boldsymbol{y} - \boldsymbol{\pi}_i) \tag{5}$$

Equation (2.3) is called the logistic regression function, which shows the relationship between predictor variables and probability is not linear, so to get a linear relationship a transformation is often called a logit transformation. The logit form of $\pi(x)$ is expressed as $g(x)$, i.e.:

$$logit\,[\pi(x)] = g(x) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \quad (6)$$

Equation (2.6) is a form of logistic regression model relationship function called multiple logistic regression model (Hosmer and Lemeshow, 2000). To obtain estimates from logistic regression parameters, it can be done using Maximum Likelihood Estimation (MLE).

The MLE method is used to estimate the parameters in logistic regression and basically, the maximum likelihood method provides an estimate of β by maximizing its likelihood function (Hosmer & Lemeshow, 1989). Mathematically the likelihood function (xi, yi) can be expressed:

$$f(y_i) = \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \quad (7)$$

In general, the likelihood function is defined as the joint probability function of the random variable formed by the sample. Especially for a sample of size n with its observations (y_1, y_2, …, y_n) corresponds to a random variable (Y_1, Y_2, …, Y_n). As long as Y_i s considered to be independent, the co-probability density function is as follows:

$$g(Y_1, Y_2, \ldots, Y_n) = \prod_{i=1}^{n} f(Y_i) \quad (8)$$

2.3 Ensemble

From one dataset, many prediction models can be obtained either using different techniques or using similar algorithms. Each model then produces predictions that can differ from one another. The ensemble learning approach combines these various predictions into one final prediction. Ensemble techniques that rely on variations from the random forest and boosting approaches can provide predictions with excellent accuracy. Random forest works by making ensemble composing models in such a way that various possibilities can be optimally accommodated while boosting works iteratively so that unpredictable cases are no longer a problem.

The ensemble method can reduce classification errors effectively, and is believed to have good performance compared to using a single classifier. The ensemble method is an algorithm in Machine Learning where this algorithm combines several models to achieve a higher generalization performance than a single model can do (Peter, 2014). The main idea of the ensemble method is to combine several sets of models that solve the same problem to get more accurate model (Aziz, 2020).

2.4 Ensemble Boosting

Boosting is designed for problems related to classification and is applied to weak classifiers. Boosting is a common and effective method for building accurate classifiers by combining weak classifiers. The use of boosting is preferred because it focuses on misclassified problems and has a tendency to increase in accuracy higher than bagging. One of the popular algorithms of the boosting method is the AdaBoost or Adaptive Boost algorithm. The focus of this method is to generate a series of base classifiers.

2.5 AdaBoost

The AdaBoost concept emerged from Kearns and Valiant's question in 1988 whether weak learning could be upgraded to a strong one. The answer to this question was then answered by Schapire by building a boosting algorithm for the first time. Furthermore, this algorithm was further developed by Freund and Schapire by proposing the Adaptive Boosting concept known as AdaBoost. AdaBoost and its variants have been successfully applied to several fields (domains) because of their strong theoretical basis, accurate predictions great simplicity.

AdaBoost trains basic classifiers sequentially (iteratively) at each iteration, basic classifications are trained using training data with weight coefficients that depend on the performance of the classifier in the previous iteration to give greater weight to misclassified data. If classifiers have been trained as much as desired, then all classifiers are combined to form a final decision on the model that shows the best performance (Kégl & Busa, 2009).

The AdaBoost algorithm steps:

1. Initialize the training data weights $w_i = \frac{1}{N}, i = 1,2,\ldots,N$.
2. Repeating as many as $m = 1,2,\ldots,M$:
   a. Pair classifiers to get an estimate of the class probability $p_m(x) = \hat{p}_w(y = 1|x) \in [0,1]$, use weight w_i on training data.
   b. Specify $f_m(x) \leftarrow \frac{1}{2} log \frac{p_m(x)}{(1-p_m(x))} \in R$.
   c. Specify $w_i \leftarrow \exp[-y_i f_m(x_i)], 1 = 1,2,\ldots,N$ and renormalize until $\sum_i w_i = 1$
   d. Classification output in the form $[\sum_{m=1}^{M} f_m(x)]$

The Adaboost algorithm is an appropriate estimation procedure for adjusting logistic regression models. The procedure optimizes the exponential criteria which are up to the second-order equivalent to the binomial log-likelihood criterion.

2.6 Performance Evaluation

Evaluation of the performance of the classification method can be seen from the level of classification errors. To calculate the misclassification value, a confusion matrix can be used. A confusion matrix is called a contingency table. The classification error can be determined through the classification table. The classification table is a contingency table ($k \times k$) based on empirical data from the response variables.

Table 1. Classification Table

Observation	Prediction	
	y_1	y_2
y_1	n_{11}	n_{12}
y_2	n_{21}	n_{22}

Note:

n_{11} : The multitude of subjects from y_1 proper classification as y_1

n_{12} : The multitude of subjects from y_1 misclassified as y_2

n_{21} : The multitude of subjects from y_2 misclassified as y_1

n_{22} : The multitude of subjects from y_2 proper classification as y_2

So that the formula for the overall misclassification of the APER value is obtained:

$$APER = \left(\frac{n_{12} + n_{21}}{n}\right) \times 100\% \quad (9)$$

Then, to obtain the correct classification value the formula: $100 - APER$ is used. In addition to $APER$, there are several evaluations of the accuracy of the classification that is popularly used, namely accuracy, sensitivity, and specificity.

$$Accuration = \frac{n_{11} + n_{22}}{n_{11} + n_{12} + n_{21} + n_{22}} \times 100\%$$
$$Sensitivity = \frac{n_{11}}{n_{11} + n_{21}} \times 100\%$$
$$Spesificity = \frac{n_{22}}{n_{12} + n_{22}} \times 100\% \quad (10)$$

2.7 Credit Scoring

Credit scoring is a method used to evaluate credit risk in terms of loan applications from consumers. This method is used to classify consumers who apply for credit into good or bad groups. Credit scoring attempts to categorize the diversity of characteristics of consumers who request credit based on errors and omissions of obligations. This method produces a calculation that can be used by the credit service company to classify the requirements of consumers applying for credit to credit risk.

The credit scoring model is formed through a series of statistical processes that can be used to predict new data. The process of applying a model that has been formed is different from the process informing or making a model. In particular, a credit scoring model that is formed can be used for a long time to calculate or predict new data. During the process of forming a credit scoring model, information from consumers in the form of data is then processed with the help of statistical software. In the end, a model will be produced that has an output in the form of decisions for consumers.

3 Results and Discussion

3.1 Research Method

The research approach used in literature study and quantitative descriptive approach. A literature study is a research method by collecting library materials or theories from various references as a reference for researchers in conducting and solving research problems. Meanwhile, the quantitative descriptive method is a data analysis method used to examine a particular population or sample with a quantitative data analysis or statistics. Quantitative research emphasizes the analysis of numerical data, produces conclusions that will clarify the description of the object, and the significant relationship between the variables studied.

There are two types of data used to achieve the objectives of this study, namely secondary data of 100 non-subsidized home ownership creditors of Bank X and simulation data. The data simulation was carried out by generating 500 and 1000 data based on the characteristics of the secondary data owned. The research variable is an attribute, nature, or value of people, objects, or activities that have certain variations that are determined by the researcher to study and draw conclusions (Sugiyono, 2002). The variables used in this study consisted of the response variable, namely the attitude of the

creditor (Y) which was denoted by 0 for creditor with current status, and 1 for creditor with non-current status, while the predictor variables were obtained from the scorecard owned by Bank X Indonesia as follows:

X1: Guarantee Documents (1 = Freehold Title / Building use rights certificate, 2 = Building use rights certificate, 3 = Freehold Title)
X2: Length of Residence (Years)
X3: City Size (Scale 1-15)
X4: Education (Years)
X5: Age (Years)
X6: Collectability status (1 = No Indonesia Bank Checking, 2 = Indonesia Bank Checking Col 1, 3 = Indonesia Bank Checking Col 2)
X7: Marital status (1 = not yet married, 2 = divorced, 3 = married)
X8: Joint Income
X9: Form of business entity, 1 = commanditaire venootschap, 2 = Ltd. Non Plc, 3 = Ltd Plc, 4 = Others)
X10: Credit Period (Years)
X11: RPA (Installment Income Ratio)
X12: Occupation (1 = state-owned enterprises / regional owned enterprises, 2 = Entrepreneur, 3 = government employees, 4 = Private / Professional Peg, 5 = Others)
X13: Work Experience (Years)
X14: Number of dependents (Person)
X15: Ownership of Savings (1 = do not have, 2 = have another bank, 3 = have)
X16: Loan To Value

Before entering the data analysis process, the first step in this analysis is to conduct a simulation study with the following stages.

1. Estimating the parameters of the logistic regression model on secondary data so as to get the parameter estimation results which can be used as the initial coefficient in the simulation method.
2. Determine the initial coefficient of the binary logistic regression model
3. Create an **X** matrix (predictor variable) with the first column being 1 and the other columns being the values of the sixteen predictor variables in the original data
4. Generating response variables generated from the Bernoulli distribution with the probability density function as follows:

$$f(y_i)=\pi(x_i)^{y_i}\left(1-\pi(x_i)\right)^{1-y_i},\ y_i=0,1 \qquad (11)$$

with π is the probability of a successful event. The Bernoulli distribution parameter is shown by Equation (3.1) which is used to generate the response variable as many as the number of observations in the secondary data.

$$\pi(x)=\frac{\exp\left(\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_px_p\right)}{1+\exp\left(\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_px_p\right)}, \qquad (12)$$

5. Form a data frame with the name A consisting of predictor variables from secondary data and new response variables from the results of the generation in step 4.
6. Take observations from the data frame generated in step 5 as many as the sample size (n = 10, 100 and 1,000) by:
   a. Divide the data into 2 parts, namely the data with the minority class response variable (label 0) and the data with the majority class response variable (label 1).
   b. Taking a random sample on the data with the minority class response variable (label 0) as much as the proportion level used (p = 20%) times the sample size (n = 10, 100 and 1,000).
   c. Taking a random sample on the data with the majority class response variable (label 1) as much as the proportion level used (1-p = 80%) times the sample size (n = 10, 100 and 1,000).
   d. Combine random samples generated points b and c which will be used as simulation data.
7. Forming a new data frame with the name B which is a duplicate of the data frame A but one of the variables is changed to a new variable which is the result of a linear combination of certain variables.
8. Repeat step 6 on data frame B.

The steps in this research are as follows:

1. Divide the data into two, namely training data and testing data with a ratio of 80%: 20% based on the Pareto principle.
2. Data processing using binary logistic regression analysis. The steps at this stage are as follows:
3. Data processing using Logistic AdaBoost analysis. The steps at this stage are as follows
4. Determining the best model by comparing the results of binary logistic regression analysis and AdaBoost regression analysis, through the accuracy of the classification model of the two analyzes.

3.2 Regression Logistics Biner Analysis

Before the analysis is carried out, the initial step that needs to be done is to divide the data into two groups, namely training data and testing data. The distribution of training and testing data is 80%: 20%. Furthermore, the partial parameter significance test was carried out by testing

individually for each predictor variable. A partial test was conducted to determine the effect of each predictor variable on the response variable. The results of the parameter estimation and partial significance can be seen in table 2.

Table 2. Estimated parameters and partial tests

Variable	Estimation	p-value	Decision
Const	-2.85e+02	0.99	Accept H_0
$X_{1(2)}$	7.27e-01	0.66	Accept H_0
$X_{1(3)}$	-1.35e+00	0.54	Accept H_0
X_2	1.67e-02	0.80	Accept H_0
X_3	5.74e-02	0.79	Accept H_0
X_4	1.76e+01	0.99	Accept H_0
X_5	-1.22e-01	0.18	Accept H_0
$X_{6(2)}$	5.83e-01	0.64	Accept H_0
$X_{6(3)}$	3.22e+00	0.11	Accept H_0
$X_{7(2)}$	-1.18e+00	0.40	Accept H_0
$X_{7(3)}$	-1.40e+00	0.41	Accept H_0
$X_{8(1)}$	-1.26e+00	0.46	Accept H_0
$X_{9(2)}$	-9.85e-01	0.61	Accept H_0
$X_{9(3)}$	1.12e+00	0.52	Accept H_0
X_{10}	-2.76e-02	0.76	Accept H_0
X_{11}	-1.04e+00	0.06 .	Reject H_0
$X_{12(2)}$	4.21e+00	0.04 *	Reject H_0
$X_{12(3)}$	1.63e+00	0.34	Accept H_0
$X_{12(4)}$	-5.03e-02	0.98	Accept H_0
$X_{12(5)}$	-1.94e+01	0.99	Accept H_0
X_{13}	1.00e-03	0.93	Accept H_0
X_{14}	4.65e-01	0.26	Accept H_0
X_{152}	-1.77e+01	0.99	Accept H_0
X_{16}	6.55e-02	0.16	Accept H_0

Based on Table 2, it can be seen that only the variables X11 and X12 have a partially significant effect on Credit Collectability. Next, a simultaneous test was carried out to determine the significance of the parameters on the response variable simultaneously or as a whole. In this test, the G test or the maximum likelihood ratio test is used. The result shows the G value of 45,854 with degrees of freedom (df = 17). Taking advantage of the opportunity $\chi^2_{(0.05;17)}$ which is equal to 27.5871, then the value of G (45.854) > $\chi^2_{(0.05;17)}$ (27.5871) so that it can be decided to reject H0. At the 5% real level, it is concluded that the independent variables simultaneously have a significant effect on Creditor Collectability.

Furthermore, the binary logistic regression analysis was carried out again by excluding variables that did not have a significant effect and the parameter estimation results were obtained at table 3.

Table 3. Estimated parameters of significant variables

Var	Est	p-value	Decision
Cont	-0.89	0.24	Accept H_0
X_{11}	-0.57	0.08 .	Reject H_0
X_{122}	2.88	0.00 **	Accept H_0
X_{123}	0.43	0.61	Accept H_0
X_{124}	-0.06	0.94	Accept H_0
X_{125}	-15.99	0.99	Accept H_0

$$g(x) = -0.892 - 0.578X_{11} + 2.887X_{12(2)} + 0.431X_{12(3)} - 0.069X_{12(4)} - 15.996X_{12(5)}$$

Then the binary logistic regression model is obtained as follows:

$$\pi(x_i) = \frac{e^{g(x)}}{1 + e^{g(x)}}$$

$$\pi(x_i) = \frac{e^{(-0.892 - 0.578X_{11} + 2.887X_{12(2)} + 0.431X_{12(3)} - 0.069X_{12(4)} - 15.996X_{12(5)})}}{1 + e^{(-0.892 - 0.578X_{11} + 2.887X_{12(2)} + 0.431X_{12(3)} - 0.069X_{12(4)} - 15.996X_{12(5)})}}$$

After the model is formed, the model suitability test is then carried out to determine whether there is a significant difference between the results of the observations and the possible value of the prediction of the model with the Hosmer Test, and a p-value of 0.92 is obtained. So it can be concluded that there is no significant difference between the results of the observations and the possibility that the prediction results of the model or model obtained are appropriate. The accuracy of model classification is calculated using data testing. Table 4 below is the prediction result of the credit collectability model.

Table 4. Contingency table logistic regression analysis classification

Observation	Classification	
	PL	Non PL
PL	16	3
Non PL	1	0

Based on table 4, it is found that the number of current creditors who have been successfully classified as current creditors is as many as 16 creditors. There are 3 current creditors but classified as non-current. In addition, there is 1 creditor who pays non-currently but is classified as current.

After obtaining the predictions in Table 5, then the classification accuracy value can be calculated as follows in table 5.

Table 5. Table of classification accuracy of logistic regression analysis

Performance Evaluation	Percentage
Accuracy	80.0
Sensitivity	84.2
Specificity	0.0

Based on the classification accuracy table, it can be seen that binary logistic regression can predict observations accurately or accurately with a percentage of 80.0%. The creditors who actually paid smoothly and were successfully predicted to be current was 84.2%. Creditors paying non-current and non-current predictions are 0%. So, it can also be captured that the logistic regression model is not able to capture and classify the existence of creditors who pay non-currently to be classified as non-current.

3.3 AdaBoost Logistic Regression Analysis

Adaboost is a method that combines classifiers iteratively is made from weighted training data, with weights adjusted adaptively at each step to give increased weight to cases that had misclassification in the previous step. The classifier used in the Adaboost method is binary logistic regression.

This method begins with the initial weight of the training data, where each object will be given the same weight (wt (i)). If the training data consists of N objects, then the initial weight on each object is 1 / N, the training data used in this study is

Scenario 1: 20% *100 = 20 (secondary data)
Scenario 2: 20% * 10 = 2 (simulation data, representing high-dimensional data where the variable is larger than the sample),
Scenario 3: 20% * 500 = 100 (simulation data, representing medium-sized simulation data), and
Scenario 4: 20% * 1,000 = 200 (simulation data, representing big data based on volume criteria),

so the initial weight on each object is 0.5, 0.01, and 0.005. Then resampling the training data with returns. The next step is to classify using logistic regression analysis using training data that has been resampled.

The relationship between logistic regression and weighting voting (αt) at AdaBoost is that each model is calculated to calculate its classification error (εt), where the greater the classification error (εt) approaches the value 0.5 (guessing misclassification), it will decrease the weighting voting value αt (at εt = 0.5 value αt = 0). So that when the final classifier calculation H (xi) is carried out, the classification which has a classifier error close to 0.5 (εt = 0.5) will be given a small weighting vote. The iteration step used in this study is 1,000 iterations, this is based on the opinion of Mease and Wyner (2008) who say that AdaBoost must be run for a long time at least as many as 1,000 steps, to obtain an increasingly convergent error rate.

The result of this analysis is the creditors collectability classification. To find out how precise this method is in clarifying creditors, it is necessary to calculate the accuracy of the classification obtained from the testing data shown.

a. Scenario 1: n=100 (secondary data)

Furthermore, the classification of creditors with a large number of samples is 10. The first scenario is expected to represent the condition of high-dimensional data, namely data that represents a variable that is larger than the number of independent variables. The results of creditor classification are shown in Table 6.

Table 6. Contingency table on simulation data with n=10

Observation	Classification	
	PL	Non PL
PL	16	1
Non PL	1	2

Based on table 6, it was found that the number of current creditors who were successfully classified as current creditors was 16 creditors. There is 1 creditor which is actually current but is classified as non-current. There is 1 creditor who pays non-currently but is classified as current. And there are 2 creditors who are not smooth and are categorized as non-current.

After obtaining the classification results in table 6, it can be calculated the level of classification accuracy presented in table 7.

Table 7. Table of classification accuracy of logistic regression analysis when n=10

Performance Evaluation	Percentage
Accuracy	90.0
Sensitivity	94.1
Specificity	66.7

Based on table 7, it can be seen that binary logistic regression with ensemble boosting can predict observations accurately or accurately with a percentage of 90.0%. Creditors who actually paid

smoothly and were successfully predicted to be current were 94.1%. Creditors who pay non-smoothly and are predicted to be non-current are 66.7%.

b. Scenario 2: n=10

Furthermore, the classification of creditors with a large number of samples is 10. The first scenario is expected to represent the condition of high-dimensional data, namely data that represents a variable that is larger than the number of independent variables. The results of creditor classification are shown in table 8.

Table 8. Contingency table on simulation data with n=10

Observation	Classification	
	PL	Non PL
PL	6	1
Non PL	1	2

Based on table 8, it was found that the number of current creditors who were successfully classified as current creditors was 6 creditors. There is 1 creditor which is actually current but is classified as non-current. There is 1 creditor who pays non-currently but is classified as current. And there are 2 creditors who are not smooth and are categorized as non-current.

After obtaining the classification results in table 6, it can be calculated the level of classification accuracy presented in table 9.

Table 9. Table of classification accuracy of logistic regression analysis when n=10

Performance Evaluation	Percentage
Accuracy	80.0
Sensitivity	85.7
Specificity	67.0

Based on table 9, it can be seen that binary logistic regression with ensemble boosting can predict observations accurately or accurately with a percentage of 80.0%. Creditors who actually paid smoothly and were successfully predicted to be current were 85.7%. Creditors who pay non-smoothly and are predicted to be non-current are 67.0%.

c. Scenario 3: n=500

The result of this analysis is the creditors collectability classification. To find out how precise this method is in clarifying creditors, it is necessary to calculate the accuracy of the classification obtained from the testing data shown in table 10.

Table 10. Contingency table on simulation data with n=500

Observation	Classification	
	PL	Non PL
PL	69	0
Non PL	1	30

Based on table 10, it was found that the number of current creditors who were successfully classified as current creditors was 69 creditors. There is 1 creditor which is actually current but is classified as non-current. There are no creditors who pay non-currently but are classified as current. And there are 30 non-current creditors and are categorized as non-current.

After obtaining the predictions in Table 8, then the classification accuracy value can be calculated as table 11:

Table 11. Table of classification accuracy of logistic regression analysis when n=500

Performance Evaluation	Percentage
Accuracy	99.0
Sensitivity	98.5
Specificity	100.0

Based on table 11, it can be seen that binary logistic regression with ensemble boosting can predict observations accurately or accurately with a percentage of 99.0%. Creditors who actually paid smoothly and were successfully predicted to be current were 98.5.0%. Creditors who pay non-smoothly and are predicted to be non-current are 100.0%.

d. Scenario 4: n=1,000

The result of this analysis is the creditors collectability classification. To find out how precise this method is in clarifying creditors, it is necessary to calculate the accuracy of the classification obtained from the testing data shown in table 12.

Table 12. Contingency table on simulation data with n=1,000

Observation	Classification	
	PL	Non PL
PL	170	0
Non PL	0	30

Based on table 12, it was found that the number of current creditors who were successfully classified as current creditors was 170 creditors. There are 30 non-current creditors and are categorized as non-current. And there are no misclassified creditors.

After obtaining the predictions in Table 10, then the classification accuracy value can be calculated as table 13:

Table 13. Table of classification accuracy of logistic regression analysis

Performance Evaluation	Percentage
Accuracy	100.0
Sensitivity	100.0
Specificity	100.0

Based on table 13, it can be seen that binary logistic regression with ensemble boosting can predict observations accurately or accurately with a percentage of 100.0%. Creditors who actually paid smoothly and were successfully predicted to be current were 100.0%. Creditors who pay non-smoothly and are predicted to be non-current are 100.0%.

3.4 Comparison of Performance Evaluation

The results can then be juxtaposed for comparison so that it can be seen which method is more accurate and results in sensitivity and specificity of creditor classification. A comparison of these methods is presented in Table 14.

Table 14. Comparison of classification accuracy

Performance Evaluation	Logistic Regression	AdaBoost Logistic Regression Analysis			
		Secondary	n = 10	n = 500	n = 1000
Acc	80.0	90.0	80.0	99.0	100.0
Sens	84.2	94.1	85.7	98.5	100.0
Spec	0.0	66.7	67.0	100.0	100.0

Based on table 14, it can be seen that in the same data, namely secondary data, the logistic regression ensemble produces a better level of accuracy, sensitivity, and specificity compared to classical logistic regression analysis. Where accuracy increased from 80% to 90%. Sensitivity increased from 84.2% to 94.1%, and specificity from 0% to 66.7%. In practice, this will have a very big effect because a misclassification will be very detrimental to both parties, both the bank and the creditors. When a creditor whose collectability is current is classified as non-current, it will cause the creditor to not obtain credit and lead to disappointment and reduced loyalty. This is also detrimental to the bank because it loses potential creditors. In other conditions, when the actual creditor has poor collectability then the classification results show that he has current collectability will cause losses to the bank later. Meanwhile, from the various scenarios applied to perform data simulation, namely at very small n even smaller than the number of independent variables, the logistic regression ensemble is able to predict customers well with an accuracy rate of 80%, a sensitivity level of 85.7% and a specificity of 67%. . While the effect of increasing the sample size can be seen that the larger the sample size, the higher the accuracy, sensitivity, and specificity of the classification model.

4 Conclusions and Recommendations

Credit is the largest asset managed by a bank and is also the most dominant contributor to bank income. Therefore, every bank applies the principle of prudence in lending. Through logistic regression, it is known that the variables that have the most influence on creditor collectibility are X11 and X12, namely the Ratio of Installment Income and Type of Creditor's Work. It is also known that the ensemble boosting logistic regression analysis resulted in more accurate classification of up to 100% in classifying creditors. The sensitivity and specificity generated by the machine learning enhancement also show a percentage of 100%, which means that no prediction errors occur. Even in high-dimensional dataset conditions, the logistical ensemble is still able to predict quite well.

Based on the results obtained, suggestions that can be given for further research are to compare the classification of creditors using other boosting methods or combine AdaBoost with other classification methods other than logistic regression analysis which can be more accurate in classification. Another thing that can be done is to further examine the credit imbalance and overcome it using the ensemble method. In addition, banks must use the logistic ensemble classification method in classifying the collectibility of their creditors.

5 Benefits of Research

The research benefits that can be provided from this research are as follows:

1. For the bank, this research can help to reduce credit risk and help select prospective creditor who meet the requirements and do not receive credit.

2. For researchers, this research is useful to increase knowledge and skills regarding the credit scoring classification process (Credit Scoring).
3. For educational institutions of Masters in Statistics, it can be used as a scientific reference in research for the development of the credit scoring classification process (Credit Scoring).

References:

[1] De Menezes, F. S., Liska, G. R., Cirillo, M. A., & Vivanco, M. J. (2017). Data classification with binary response through the Boosting algorithm and logistic regression. *Expert Systems with Applications*, *69*, 62-73.

[2] Fernandes, A. (2020). Comparison of Parameter Estimator Efficiency Levels of Path Analysis with Bootstrap and Jack Knife (Delete-5) Resampling Methods on Simulation Data. *Jurnal Matematika, Statistika dan Komputasi*, *16*(3), 353-364.

[3] Fernandes, A. A. R. (2019). The Estimation Function Approach Smoothing Spline Regression Analysis for Longitudinal Data. In *IOP Conference Series: Materials Science and Engineering* (Vol. 546, No. 5, p. 052064). IOP Publishing.

[4] Fernandes, A. A. R. (2018). *Metodologi Penelitian Kuantitatif Perspektif Sistem: Mengungkap Novelty dan Memenuhi Validitas Penelitian*. Universitas Brawijaya Press.

[5] Friedman, J., Hastie, T., & Tibshirani, R. (2000). Special invited paper. additive logistic regression: A statistical view of boosting. *Annals of statistics*, 337-374.

[6] Hosmer, D. W., & Lemesbow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in statistics-Theory and Methods*, *9*(10), 1043-1069.

[7] Kirono, I., Armanu, A., Hadiwidjojo, D., & Solimun, S. (2019). Logistics performance collaboration strategy and information sharing with logistics capability as mediator variable (study in Gafeksi East Java Indonesia). *International Journal of Quality & Reliability Management*.

[8] Lewis, R.J. (2000). *An Introduction to Classification And Regression Tree (CART) Analysis*. Annual Meeting of the Society For Academic Emergency Medicine in San Fransisco. California: Department of Emergency Medicine.

[9] Pham, B. T., Bui, D. T., & Prakash, I. (2017). Landslide susceptibility assessment using bagging ensemble based alternating decision trees, logistic regression and J48 decision trees methods: a comparative study. *Geotechnical and Geological Engineering*, *35*(6), 2597-2611.

[10] Sugiyono, D. (2002). Metode Penelitian Bisnis: Bandung, CV.

[11] Xu, X., & Frank, E. (2004). Logistic regression and boosting for labeled bags of instances. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 272-281). Springer, Berlin, Heidelberg.

[12] Yohannes, Yesihac; Hoddinott, John. (1999). *Classification And Regression Tree: An Introduction*. Washington, DC: Internationl Food Policy Research Institut.