Determining the Eligibility of Providing Motorized Vehicle Loans by Using the Logistic Regression, Naive Bayes and Decision Tree (C4.5)

Harsih Rianto¹, Amrin², Rudianto³, Omar Pahlevi⁴, Paramita Kusumawardhani⁵, and Seno Sudarmono Hadi⁶

¹,³,⁴ Program Studi Sistem Informasi, Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika, Jl. Kramat Raya No. 98 Jakarta, Indonesia
² Program Studi Teknologi Komputer, Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika, Jl. Kramat Raya No. 98 Jakarta, Indonesia
⁵ Program Studi Bahasa Inggris, Fakultas Komunikasi dan Bahasa, Universitas Bina Sarana Informatika, Jl. Kramat Raya No. 98 Jakarta, Indonesia
⁶ Program Studi Manajemen, Fakultas Ekonomi dan Bisnis, Universitas Bina Sarana Informatika, Jl. Kramat Raya No. 98 Jakarta, Indonesia

E-mail: harsih.hhr@bsi.ac.id

Abstract. Evaluating in determining the eligibility of giving credit is very important. Errors in providing credit worthiness assessments can result a bad credit risk. The problem that often occurs is not the application of the system by financial parties but more on HR when making predictions about the determination of consumer credit worthiness. Research in the field of computers has been done to reduce credit risk resulting in losses to the company. In this research a comparison of Logistic Regression (LR), Naïve Bayes (NB) and Decision Tree (C4.5) algorithms is performed to predict the feasibility of granting credit. In order to produce a prediction of the feasibility of granting credit to new consumers, credit data used by the company is used. The data used in this study consists of 481 consumer records that have been classified as consumers with current credit and bad credit. After testing using the same dataset on the three algorithms by comparing the AUC and Confusion Matrix values, it was found that the appropriate algorithm to be applied to the credit worthiness dataset was Logistic Regression with an Area Under Curve (AUC) value of 0.972 and Accuracy or Confusion Matrix of 93.14%. As for the Decision Tree Algorithm (C4.5) from the test results, the AUC value is 0.926 and the Accuracy is 90.85% and the Algortima Naïve Bayes AUC value is 0.905 and the Accuracy is 82.75%.

1. Introduction

The use of motorized vehicles is now a necessity of the community. Production of two-wheeled or four-wheeled vehicles continues to increase to meet the needs of the community. The high price of motor vehicles causes people to choose credit to buy it [1]. The number of cases of bad debts occurred before the repayment period due to the analysis of the previous credit disbursement was not done correctly and appropriately [2]. The process of evaluating credit worthiness today is mostly still using the manual method, by giving forms that are filled out by prospective credit recipients [3]. So that the analysis can be made error assessing credit worthiness. To reduce the
occurrence of bad credit, it is necessary to evaluate the feasibility of granting credit with the right method. Evaluating through the data that the company already has is the right choice by using a data mining model.

Data Mining is the process of collecting, cleaning, processing, analyzing, and obtaining useful information from data [4]. The process is repeated to find the prediction results through automatic or manual methods[5]. Techniques that can be performed using data mining are estimation, association and classification [6]. Classification is currently a popular approach to predict the feasibility of providing motor vehicle loans. Many studies use classification algorithms such as Linear Regression, Logistic Regression, Naive Bayes, Neural Networks, Random Forest, Decision Tree and Support Vector Machine. [7].

In previous studie by Astuti, the results of the Accuracy Comparison of Decission Tree (C4.5) algorithm were 92.8%, K-Nearest Neighbor (K-NN) was 77.78% and Neural Network was 91.1% [2]. Then the research conducted by Amrin using the Neural Network Multilayer Perception algorithm produces an Accuracy of 96.1% [8], while the application of the Neural Network algorithm of the Radial Basis Function model produces an accuracy of 89.2% [1]. Neural Network algorithm produces a accuracy value that is good enough for the credit analysis process, but has problems with the dataset used. It is necessary to convert the attribute values that must be adjusted so that the Neural Network algorithm can run.

Comparative algorithm result obtained by two algorithms with the best value, namely Logistic Regression and Naïve Bayes [7]. Naïve Bayes is a simple probability classification model. The use of Naïve Bayes algorithm is very easy and convenient because it does not require complicated and difficult parameter estimation. Naïve Bayes presents the results of classification to users very easily and without having to have prior knowledge of classification technology [9]. Logistic Regression is a statistical classification method of probability. The advantages of Logistic Regression are that it is easy to implement, cheaper in the computation method and easy to understand the classification results [10]. However Logistic Regression also has weaknesses which are vulnerable to underfeeding problems and have low accuracy [11].

In this study the training data used is a motor vehicle credit dataset from PT Buana Kredit Sejahtera which is engaged in financing. This motorized vehicle credit dataset consists of 481 records with good credit status and bad credit status (bed). The dataset will be used in the method of determining credit worthiness and comparing the Regression Logistic, Naïve Bayes and Decision Tree (C4.5) algorithms to find the best solution in determining the creditworthiness of determining motor vehicle loans. The final aim of this study is to predict the feasibility of providing motor vehicle loans.

2. Method

Data mining is the process of shifting from a large database to look for patterns that were not previously known and produce new knowledge [12]. In data mining can use the classification model to get the desired model. The classification model is the process of analyzing data to help people make class predictions from the sample labels to be classified. [13][14].

2.1. Logistic Regression

Logistic regression is presented for prediction using more than one Linear Regression [15]. Logistic Regression shows the linear equation that is interconnected between several random variables, where the dependent variable is the continuous variable. Logistic Regression Model is a probability of several events, this method uses a linear function to calculate predictions on several variables [12]. Logistic Regression uses predefined variables or variables that are categorized into 2 variables. As in the prediction of success or failure, life or death, sick or not sick, and so forth.
Notation in Logic Regression where variable $X \in \mathbb{R}^{N \times d}$ is a data matrix where $N$ is the number of samples, $d$ is the number of parameters or attributes, and the variable $y$ is a binary outcome variable. The commonly used formula of Logistic Regression for each $X_i$ sample model adheres to the function of Hosmer [16] as follows:

$$E[y_i | X_i | \beta] = P_i = \frac{e^{X_i \beta}}{1 + e^{X_i \beta}}$$  \hspace{1cm} (1)

Where $\beta$ is the vector parameter assuming $X_{i0} = 1$ so it is a constant variable with a value of 1. Transformation of logistic regression is a logit which produces a formulation to get the odds value with the response variable being positive, can be seen in Formula 2. as follows:

$$n_i = \ln\left(\frac{P_i}{1-P_i}\right) = X_i \beta$$  \hspace{1cm} (2)

In the form of a matrix, the logistic formula can be formulated as:

$$n = X \beta$$  \hspace{1cm} (3)

Now with the assumption that observation of sample data use independent variables, the likelihood function can be written as follows:

$$L(\beta) = \prod_{i=1}^{l} (p_i)^{y_i} (1-p_i)^{1-y_i} = \prod_{i=1}^{l} \left(\frac{e^{X_i \beta}}{1 + e^{X_i \beta}}\right)^{y_i} \left(\frac{1}{1 + e^{X_i \beta}}\right)^{1-y_i}$$  \hspace{1cm} (4)

Furthermore, using the formula from Hosmer [16] to obtain log likelihood is written as follows:

$$\log L(\beta) = \sum_{i=1}^{l} \left(y_i \log \left(\frac{e^{X_i \beta}}{1 + e^{X_i \beta}}\right) + (1-y_i) \log \left(\frac{1}{1 + e^{X_i \beta}}\right) \right) - \frac{\lambda}{2} \|\beta\|^2$$  \hspace{1cm} (5)

$$= - \sum_{i=1}^{l} \log \left(e^{y_i X_i \beta}(1 + e^{X_i \beta})\right) - \frac{\lambda}{2} \|\beta\|^2$$  \hspace{1cm} (6)

Where there are $\frac{\lambda}{2} \|\beta\|^2$ provisions added to get better generalizations. Since the formula tightening was applied in determining log likelihood, this formula was objectively used to find Maximum Likelihood Estimation (MLE)[17], $\beta$ which maximizes log likelihood[17]. For binary results, missing functions or deviations from log likelihood functions are used from Hosmes [16] namely:

$$DEV(\beta) = -2\ln L(\beta)$$  \hspace{1cm} (7)

Minimizing deviations $DEV(\beta)$ as done by Busser et al., 1999 in Maalouf and Trafalis [18] is the same as maximizing log-likelihood by King and Zeng [19]. Deviations in non-linear is about data functions. Minimization is needed on numerical data to find Maximum Likelihood Estimate (MLE) results from $\beta$.

2.2. Naïve Bayes

The Naïve Bayes algorithm is a simple classification [14] and has a fairly good reputation when making predictions resulting in high accuracy [20]. In dealing with the classification problem the Naïve Bayes algorithm has been used extensively, in addition to the time required in the learning process requires a faster time compared to other Learning Machine Algorithms [21]. Naïve Bayes is a simple probability classifier that applies the Bayes theorem with a high independent assumption.
Naïve Bayes are statistical classifiers that can be used to predict the probability of membership of a class. Naïve Bayes is proven to have high accuracy and speed when applied to databases with large data. Bayes’ theorem uses the following equation:

\[ P(H|X) = \frac{P(X|H)P(H)}{P(X)} \] (8)

### 2.3. Decision Tree

Decision Tree is a simple structure that can be used as a classification and is also popular [14]. In the classification process, the Decision Tree uses a top-down approach, thus providing a fast and effective method for classifying data [23]. Generally, a Decision Tree is a set of rules. Decision Tree will divide the training dataset into smaller partitions until each partition has only one class.

Decision Tree will form a decision tree node, each internal node (non-leaf) represents a variable attribute value and each branch represents a state of the variable that is owned. The root node selected by the Decision Tree will be selected based on the information gain attribute with the highest value. The general form for determining information gain is as follows:

\[ \text{Info}(D) = -\sum_{i=1}^{n} P_i \log_2(P_i) \] (9)

Info (D) or total information entropy is the average amount of information needed to find the value of Ci from a Xi ∈ instance D. The purpose of the Decision Tree is to partition D repeatedly into D1, D2, D3, ..., Dn with all instances in each Di belong to the same Ci class. Next root will be taken from the attributes that will be selected, by calculating the gain value of each attribute, the highest gain value will be the first root. In weighting, the accuracy value can be written with the following entropy calculation formula:

\[ \text{Gain}(S, A) = S - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \frac{|S_i|}{|S|} \] (10)

From this formula, S is the set of cases, n is the number of partitions or class classification S and \( P_i \) is the number of sample proportions (opportunities) for class i. Entropy is a parameter to measure the level of diversity (heterogeneity) of a data set.

### 2.4. Model Evaluation and Validation

In testing the dataset used overall accuracy is generally used to evaluate the performance of classification algorithms [24]. To measure the accuracy of the model, an evaluation and validation is performed using the Confusion Matrix and Area Under the ROC (Receiver Operating Characteristic) Curve techniques. Confusion Matrix is a tool for visualizing model learning outcomes in supervised learning. Each column in the matrix is an example of a prediction class, whereas each row represents the actual event in the class [25]. Confusion Matrix contains actual and predictive information on the classification model.

Furthermore, the application of evaluation uses the Area Under Curve (AUC) to measure the results of accuracy as an indicator of the performance model being run. AUC produces two lines with positive true form as vertical line and false positive as horizontal line [26]. The ROC curve is a graph between sensitivity (true positive rate), on the Y axis with 1-specificity on the X-axis (false positive rate), this ROC curve depicts as if there was an attraction between the Y-axis and the X-axis [27].
3. Results and Discussion
Research conducted in this study uses a computer platform based on Intel Core i3-3217U @ 1.80GHz (4 CPUs), 2GB RAM, and Microsoft Windows 7 Ultimate 32-bit operating system. The software used to analyze is Rapid Miner version 9.5.001. The dataset used was 481 credit data of motorized vehicles both problematic and non-problematic. The input variables in this study consisted of thirteen variables, namely: 1) Marital Status, 2) Number of Dependents, 3) Age 4) Residence Status, 5) Home Ownership, 6) Employment, 7) Employment Status, 8) Company Status, 9) Income, 10) Advances, 11) Education, 12) Length of Stay, and 13) Housing Conditions. Table 1 shows a sample dataset used to test the algorithm to be tested.

| Merital Status | Number Of Dependents | Age  | ... | Home Ownership |
|----------------|----------------------|------|-----|----------------|
| Menikah        | 2-3                  | 21-55| ... | 5              |
| Belum Menikah  | 0                    | 21-40| ... | 5              |
| Menikah        | 2-4                  | 30-40| ... | 4              |
| Menikah        | 0                    | 30-40| ..  | 2              |
| Belum Menikah  | 0                    | 30-40| ... | 3              |
| ...            | ...                  | ...  | ... | ...            |
| ...            | ...                  | ...  | ... | ...            |
| ...            | ...                  | ...  | ... | ...            |
| Menikah        | 2-4                  | 30-40| ... | 4              |
| Menikah        | 0                    | 30-40| ..  | 2              |
| Menikah        | 2                    | 20-40| ... | 2              |

This research was conducted by testing experiments on the proposed model. Then the model is evaluated and validated to produce accuracy and AUC values. Testing uses Rapid miner with a 10-fold cross-validation operator to get accuracy and AUC results on each algorithm that is tested using a motor vehicle credit dataset. The 10-fold cross-validation process will divide the dataset in the first part into testing data and the second part up to the tenth part data into training data. Figure 1 shows the comparative model that was built. Testing conducted in this study is to use the cross validation method. Cross validation is a statistical method used to
evaluate and compare algorithms by dividing data into two segments, the first segment is used as training data and the second segment is as testing data to validate the model [11]. In cross validation the training segment and the testing segment must be crossover so that each data has a validated opportunity. The cross validation model which is divided into two segments can be seen in Figure 2.

The evaluation conducted is by Confusion Matrix and ROC Curve or Area Under Curve (AUC), Confusion Matrix contains information about the actual classification and prediction made by the classification system assuming that:

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

Thus the accuracy value of the confusion matrix can be formulated by:

Next is the evaluation stage which displays the form of the model’s performance report. Table 1 is the result of comparing the accuracy, precision, recall and AUC values of the three classification algorithms. The Logistic Regression model produces the best Accuracy value with a value of 93.14%. AUC value from Logistic Regression is also the best value with a value of 0.972. Whereas the Naive Bayes algorithm accuracy value is 82.75% which is the lowest result of the three Algorithms tested.

| Algorithms          | Accuracy | Precision | Recall   | AUC   |
|---------------------|----------|-----------|----------|-------|
| Logistic Regression | 93.14%   | 96.11%    | 93.05%   | 0.972 |
| Naive Bayes         | 82.75%   | 86.59%    | 86.12%   | 0.905 |
| Decision Tree (C4.5)| 90.85%   | 92.83%    | 92.72%   | 0.926 |

From the results of the comparison of accuracy values in Table 1, it can be seen that the most superior algorithm used for the vehicle credit dataset is the Logistic Regression Algorithm. Then a comparison test is performed between each variable obtained by using t-test. It can be seen in Table 2 the results of the t-test performance.

From the t-test results it can be seen that there are significant differences between Logistic Regression, Naïve Bayes and Decision Tree (C4.5) so that it can be said that Logical Regression has significant differences with other algorithms.
Table 3. T-test Performance

|                      | Decision Tree (C4.5) | Logistic Regression | Naive Bayes |
|----------------------|----------------------|---------------------|-------------|
| Decision Tree (C4.5) | 0.189                | 0.003               |             |
| Logistic Regression  |                      |                     | 0.000       |
| Naive Bayes          |                      |                     |             |

Figure 3 is a graph from the comparison table of the accuracy of the Logistic Regression, Naïve Bayes and Decision Tree (C4.5) algorithms which shows that the highest chart is shown by the comparative results of the Logistic Regression algorithm.

![Figure 3. 10 Fold Cross Validation](image)

4. Conclusions
The experimental results in this study obtained an accuracy value of 93.14% in the motor vehicle credit dataset with the Logistic Regression algorithm model, the results of the Naïve Bayes algorithm model with a value of 82.75%. The results of comparisons with other researchers using other classifications (Neural Network and K-Nearest Neighbor (K-NN)) show the accuracy of the Logistic Regression aimed at the best results, both in accuracy and AUC parameters.

The conclusions obtained through tests that have been done show that the Logistic Regression algorithm can analyze the risk of bad credit by making predictions that are close to the correct results. Furthermore, interventions can be made that can be used by companies in analyzing lending. This research can be further developed by combining optimization models such as Bagging, Bootstrapping, feature selection and other classification algorithms.

References
[1] A. Amrin, “Perbandingan Metode Neural Network Model Radial Basis Function Dan Multilayer Perceptron Untuk Analisa Risiko Kredit Mobil,” Paradigma, vol. XX, no. 1, pp. 31–38, 2018.
[2] P. Astuti, “Komparasi Penerapan Algoritma C45, KNN dan Neural Network Dalam Proses Kelayakan Penerimaan Kredit Kendaraan Bermotor,” Fakt. Exacta, vol. 9, no. 1, pp. 87–101, 2016.
[3] A. Amrin, “Analisa Kelayakan Pemberian Kredit Mobil Dengan Menggunakan Metode Neural Network Model Radial Basis Function,” Paradigma, vol. 19, no. 102, pp. 1410–5063, 2017.
[4] C. C. Aggarwal, Data Mining: The Textbook. 2015.
[5] M. Kantardzic, DATA MINING. Concepts, Models, Methods & Algorithms. 2011.
[6] H. Rianto and R. Wahono, “Resampling Logistic Regression untuk Penanganan Ketidakseimbangan Class pada Prediksi Cacat Software,” J. Softw. Eng., vol. 1, no. 1, pp. 46–53, 2015.
[7] T. Hall, S. Beecham, D. Bowes, D. Gray, and S. Counsell, “A Systematic Literature Review on Fault Prediction Performance in Software Engineering,” IEEE Trans. Softw. Eng., Nov. 2012.
[8] A. Amrin and I. Satriadi, “Implementasi Jaringan Syaraf Tiruan Dengan Multilayer Perceptron Untuk Analisa Pemberian Kredit,” J. Ris. Komput., vol. 5, no. 6, pp. 605–610, 2018.
[9] X. Wu and V. Kumar, The Top Ten Algorithms in Data Mining. Taylor & Francis Group, 2010.
[10] S. Canu and A. Smola, “Kernel methods and the exponential family,” Neurocomputing, Mar. 2006.
[11] P. Harrington, Machine Learning in Action. Manning Publications Co, 2012.
[12] I. H. Witten, E. Frank, and M. A. Hall, Data Mining Practical Machine Learning Tools and Techniques Third Edition. 2011.
[13] L. Rokach, “Ensemble-based classifiers,” Artif. Intell. Rev., vol. 33, no. 1–2, pp. 1–39, 2010.
[14] S. Lessmann, B. Baesens, C. Mues, and S. Pietsch, “Benchmarking classification models for software defect prediction: A proposed framework and novel findings,” IEEE Trans. Softw. Eng., vol. 34, no. 4, pp. 485–496, 2008.
[15] P. Karsmakers, K. Pelckmans, and J. a. K. Suykens, “Multi-class kernel logistic regression: a fixed-size implementation,” 2007 Int. Jt. Conf. Neural Networks, Aug. 2007.
[16] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression Third Edition. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2013.
[17] M. Maalouf and M. Siddiqi, “Weighted logistic regression for large-scale imbalanced and rare events data,” Knowledge-Based Syst., 2014.
[18] M. Maalouf and T. B. Trafalis, “Robust weighted kernel logistic regression in imbalanced and rare events data,” Comput. Stat. Data Anal., Jan. 2011.
[19] G. King and L. Zeng, “Logistic Regression in Rare,” Polit. Anal., 2001.
[20] T. Menzies, J. Greenwald, and A. Frank, “Data Mining Static Code Attributes to Learn Defect Predictors,” vol. 33, no. 11, pp. 2–14, 2007.
[21] B. Turhan and A. Bener, “Analysis of Naive Bayes’ assumptions on software fault data: An empirical study,” Data Knowl. Eng., vol. 68, no. 2, pp. 278–290, 2009.
[22] D. R. Ente, S. Arifin, Andereza, and S. A. Thamrin, “Comparison of C4.5 algorithm with naive Bayesian method in classification of Diabetes Mellitus (A case study at Hasanuddin University hospital Makassar),” J. Phys. Conf. Ser., vol. 1341, no. 9, 2019.
[23] B. Chandra and P. Paul Varghese, “Fuzzifying Gini Index based decision trees,” Expert Syst. Appl., vol. 36, no. 4, pp. 8549–8559, 2009.
[24] R. Chang, X. Mu, and L. Zhang, “Software Defect Prediction Using Non-Negative Matrix Factorization,” J. Softw., Nov. 2011.
[25] F. Gorunescu, “Data mining: Concepts, models and techniques,” Intell. Syst. Ref. Libr., vol. 12, 2011.
[26] C. Vercellis, Business Intelligence: Data Mining and Optimization for Decision Making. John Wiley & Sons, 2011.
[27] R. Dubey, J. Zhou, Y. Wang, P. M. Thompson, and J. Ye, “Analysis of sampling techniques for imbalanced data: An n=648 ADNI study,” Neuroimage, 2014.