www.ijemr.net ISSN (ONLINE): 2250-0758, ISSN (PRINT): 2394-6962

Volume-7, Issue-4, July-August 2017

International Journal of Engineering and Management Research

Page Number: 393-397

Credit Scoring Analysis using LASSO Logistic Regression and Support Vector Machine (SVM)

Putri Dina Sari[1], Muhammad Nur Aidi[2], Bagus Sartono[3]
[1,2,3]Department of Statistics Bogor Agricultural University, INDONESIA

ABSTRACT

Credit scoring analysis using logistic regression is less effective, as it requires hypothesis and assumption testings that must be fulfilled. Therefore, LASSO logistic regression analysis can be used to overcome this problem because it does not require hypothesis and assumption testings. In addition, Support Vector Machine (SVM) method is also a classification method that has good classification capability as well as can ignore all assumptions such as logistic regression. This study aims to study the factors that affect the smoothness of debtors in paying motorcycle credits and compare the goodness of both methods used. Explanatory variables that affect the smoothness of credit payments are the phone ownership, down payment, loan term, occupation, age, marital status, gender, education level, number of dependents, motorcycle type, interaction between motorcycle type with phone ownership and down payment, interaction between phone ownership with occupation, loan terms, and phone number ownership, interaction between downpayment with loan terms, installments, gender and educational level, and interaction between occupation with tyype of income. Application of LASSO logistic regression and SVM in this case have mostly the same classification accuracy. However, the results of the classification performance using SVM method is relatively stable compared with LASSO logistic regression

***Keywords*--** logistic regression, LASSO, classification, support vector machine

I. INTRODUCTION

The results of banking surveys showed that credit growth in Indonesia in 2016 reached 12% (yoy). This is higher than the realization of credit growth in 2015 (as of November 2015) at 9.8% (yoy). The optimism of credit growth was mainly driven by the increasing liquidity condition of the banks, the continuing decline in lending rates, and the expected better economic conditions [1]. According to UU no. 10 of 1998, credit is a money supply or equivalent claim based on a loan agreement or agreement between a bank and another party requiring the borrower to repay his debt after a certain period of time with interest.

The high demand for credit does not necessarily making the credit service providers have to be able to grant all existing applications. Thus, it is need to do a selection process in order to find out which debtor candidates are allowed given a loan. Credit scoring is a method to evaluate credit risk that may occur in debtors by using a credit score that will be made based on the model. Credit score is a number that will represent a debtor's eligibility to receive credit grant.

Analysis of Credit scoring using logistic regression is less effective, as it requires hypothesis and assumption testings that must be fulfilled. LASSO logistic regression analysis can be used to overcome this problem [2]. In addition, the Support Vector Machine (SVM) method is one of the classification methods that can ignore assumptions the same as that of logistic regression analysis. Moreover, SVM has a good classification capability as well as can be used for large-scale data with many explanatory variables and robust to outliers [3]

This study aims to study the factors that affects the smoothness of debtors in paying motorcycle credits and compare the goodness of both methods used, that are LASSO logistic regression analysis and Support Vector Machine (SVM) methods.

II. METHOD

The data used were secondary data obtained from a branch of a consumer finance company in the form of purchasing motorcycle credits. The motorcycle purchase data was taken for 42 months from January 2008 to June 2011. The total debtor consists of 13494 customers with 16 characteristics of each debtor on the purchase of motorcycle credits.

ISSN (ONLINE): 2250-0758, ISSN (PRINT): 2394-6962

Table 1. Variable of Credit Debtor Data

Variable	detail	coding
Y	Debtor category	[0] *Good* [1] *Bad*
X_1	Motorcycle type	[1] Lower Class Underbone [2] Upper Class Underbone [3] scooters [4] Sport
X_2	Phone number ownership	[0] Not fixed [1] Fixed
X_3	Phone ownership	[0] Not fixed [1] Fixed
X_4	Down payment (percent)	[1] < 5% [2] 5% - 10% [3] 10% - 15% [4] 15% - 20%
X_5	Loan terms (month)	[1] < 13 [2] 13 - 24 [3] 25 - 36 [4] >36
X_6	Gender	[0] Male [1] Female
X_7	Income (million)	[1] < 2 [2] 2 – 3,5 [3] ≥ 3,5
X_8	Total Income (million)	[1] < 2 [2] 2 – 3,5 [3] ≥ 3,5
X_9	Monthly installment (million)	[1] < 0,5 [2] 0,5 – 0,749 [3] 0,75 – 1,249 [4] ≥ 1,25
X_{10}	Home ownership	[1] Own House [2] Family House [3] Rent / contract house [4] Home Office / Credit / Apartment, etc.
X_{11}	Marital Satus	[1] Married [2] Not Married [3] Widower / Widow
X_{12}	Number of dependents (person)	[1] 0 [2] 1 [3] 2 [4] 3 [5] 4 [6] 5 [7] > 5
X_{13}	Level of education	[1] No school / elementary school [2] Junior High School [3] Senior High school [4] Diploma [5] Bachelor Degree [6] Postgraduate
X_{14}	Occupation	[1] Private Employees [2] Farmers [3] Traders [4] Civil Servants [5] Teachers / educators [6] Services [7] Labor, etc.
X_{15}	Income Type	[0] Not fixed [1] Fixed
X_{16}	Age (years)	[1] < 21 [2] 21- 30 [3] 31- 40 [4] 41 - 50 [5] > 50

Procedur Data Analysis

1. Exploration of response variable and explanatory variables.
2. Calculating the value of Weight of Envidance (WoE).

In this section we calculate the WoE value for the category of each explanatory variable. A list of explanatory variables is shown in Table 1. The calculation of WoE values is performed using the equation:

$$WoE(i) = log\left(\frac{f_G(i)}{f_B(i)}\right)$$

where

$$f_G(i) = 100\left(\frac{n_{Gi}}{n_G}\right) = \text{Percentage of good individuals in the } i^{th} \text{ category}$$

$$f_B(i) = 100\left(\frac{n_{Bi}}{n_B}\right) = \text{Percentage of bad individuals in the } i^{th} \text{ category}$$

3. Grouping of data into 2 parts namely training data (70%) and test data (30%).
4. Conducting credit scoring analysis using LASSO logistic regression.

Stages of Data Analysis are:

a. Finding the value of *t*, with $0 \leq t$, where *t* is tuning parameter with selected value from cross validation
b. Conducting iteration process so as to obtain a convergent t value, so as to obtain the alleged parameter β.
c. Make LASSO logistic regression model
d. Testing goodness of the model by calculating the value of accuracy, sensitivity, and specificity.

5. Conducting credit scoring analysis using Support Vector Machine (SVM) [4]

Stages of the analysis are:

a. Choosing the kernel functions to be used, that are linear, radial, polynomial and sigmoid.
b. Mapping training data from input space into feature space using kernel functions.
c. Estimating model parameters for each kernel function using the quadratic programming.
d. Building a classification model for the separators between the two classes.

ISSN (ONLINE): 2250-0758, ISSN (PRINT): 2394-6962

e. Conducting goodness testing of classification model by calculating the value of accuracy, sensitivity, and specificity.
6. Comparing the level of accuracy obtained from logistic regression analysis of LASSO and SVM.
7. Selection of explanatory variables that affect in credit scoring analysis using LASSO logistic regression method

III. OUR APPROACH

Data Description

The description of the data in Figure 1 illustrates the characteristics of the debtor. Of the 13494 debtors observed in the study, 58% of debtors are able to repay the loan installments or were classified into the good debtor while 42% of debtors had defaulted or classified into bad debtor

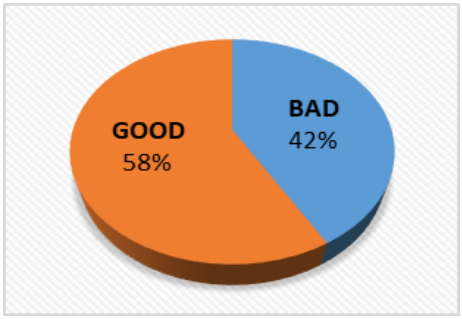

Figure 1. Debtor Characteristics based on Payment Status

Description of the data for each explanatory variable is shown in Figure 2 to Figure 4. The biggest percentage of debtors default for motorcycle type variable is obtained by lower class underbone (*Bebek*) that is equal to 26% (Figure 2a). Based on Figure 2b, the highest percentage of debtors suffering default is the debtor who has a not fixed phone number (39%). This indicates that the risk of default is more susceptible to borrowers with not fixed phone numbers. The majority of debtors already have fixed telephone are 66% of which 34% of them suffering default (Figure 2c). Subsequently, in Figure 3d the highest percentage of debtors suffering default is those who pay down payment less than 5%. From the Figure, it can be concluded that the smaller the down payment paid the greater the percentage of default.

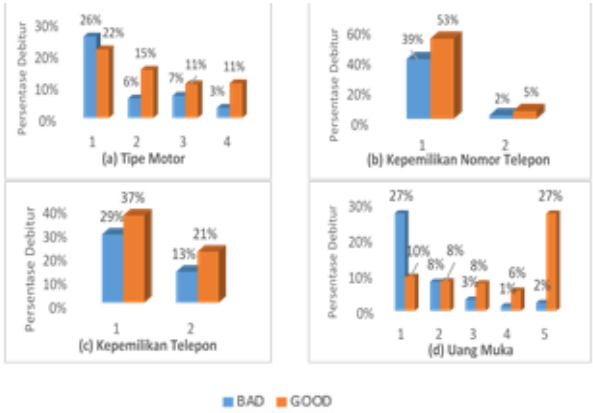

Figure 2. Percentage of debtors with paid and default on explanatory variables (a) motorcycle type, (b) ownership of phone number, (c) ownership of phone and (d) down payment

Furthermore, the characteristics of debtors are based on the status of loan repayment term. Based on Figure 3a the highest percentage of default debtor is debtor with loan term ranging from 25 to 36 months (35%). Based on the bar chart, it can be concluded that the greater the loan term, the greater the percentage of default. The majority of debtors who perform motorcycle credits are male, which is 77% (Figure 3b). In terms of income and total revenue, the greater the debtor's income the less risk of default (Figure 3c and 3d). Based on Figure 3e, the majority of debtors prefer to choose the number of installments between Rp.500.000 to Rp.750.000 per month that is amounted to 54%. The smallest proportion of default is obtained by the debtors with the number of installments between Rp.750.000 to Rp.1.500.000 per month. 62% of total debtors already have their own home, the largest proportion of default is debtors with the home ownership status of rental house or contract (Figure 3f).

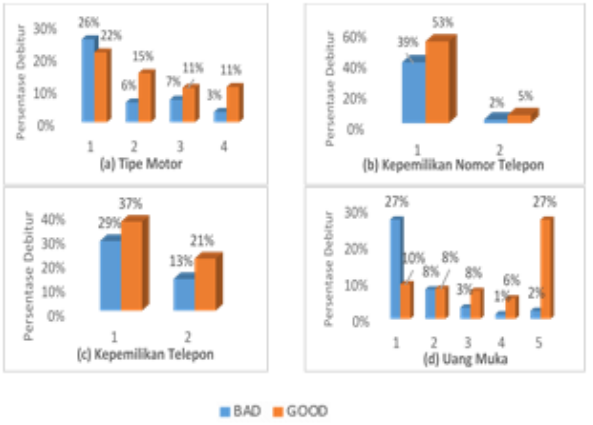

Figure 3. Percentage of debtors with repayment and default on explanatory variables (a) loan term, (b) gender, (c) income, (d) total revenue, (e) installments and (f) home ownership

ISSN (ONLINE): 2250-0758, ISSN (PRINT): 2394-6962

Marital status of 66% debtors is married, 35% of them have debt defaulted. Meanwhile, the smallest proportion of default is found in unmarried debtors (Figure 4a). Based on Figure 4b, the largest proportion of default in the dependent category occurs on the debtor with the number of dependents more than 5 persons. This is due to the larger expenditures for debtors with a total dependence on more than 5 persons compared to other debtors. The majority of debtors have an education level of senior high school, that is reached 41% (Figure 4c). Based on Figure 4d, most of the debtors have jobs as private employees and farmers that are 22% respectively, while the largest Proportion of default occurs in debtors who have jobs as traders. This is might be caused by the income from traders is not fixed for each month. On Figure 4e, the majority of debtors have a not fixed income, that is equal to 58%. The highest percentage of default is in the type of not fixed income that is reached 26%. Furthermore, Figure 4f shows that the proportion of debtors with the greatest default risk is the debtors with age between 21 to 30 years that is equal to 14%.

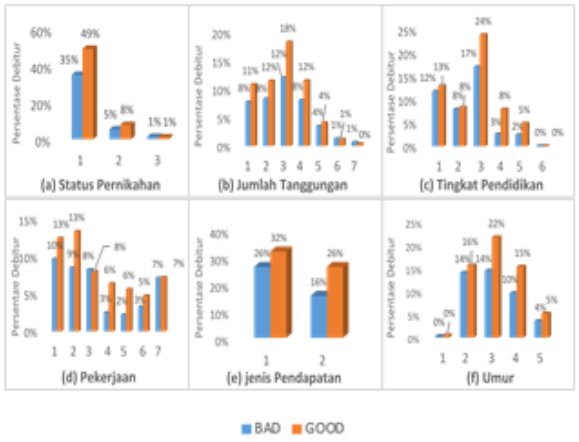

Figure 4. Percentage of debtors with repayment and default on explanatory variables (a) marital status, (b) number of dependents, (c) educational level, (d) occupation, (e) income type and (f) age

LASSO Logistic Regression

The classification using LASSO logistic regression method is done by modeling the training data (70%) to build the LASSO logistics model while the test data (30%) is used to predict the model that has been formed and then evaluated by calculating the goodness of the model. In building the LASSO logistics model is done in two stages. The first stage is to build LASSO logistic regression model with all explanatory variables. The second stage is to build LASSO logistic regression model with all explanatory variables as well as to include two interactions on each explanatory variables.

The result of LASSO logistic regression analysis at two stages is shown in Figure 5. The result of LASSO logistic regression prediction in the second stage is better than that of the first stage. In the second stage, the accuracy value obtained is 79.20%, indicates that a predicted bad debtor, in fact, is a bad debtor, in parallel a predictable good debtor in the reality is also a good debtor, the prediction could be done appropriately that equal to 79.20%. Then the sensitivity value of 72.62% indicates that a predicted bad debtor and the reality of a bad debtor is exactly classified at 72.62%. Furthermore, the specificity value of 85.02% indicates that a predictable good debtor and the reality is classified appropriately that equal to 85.02%.

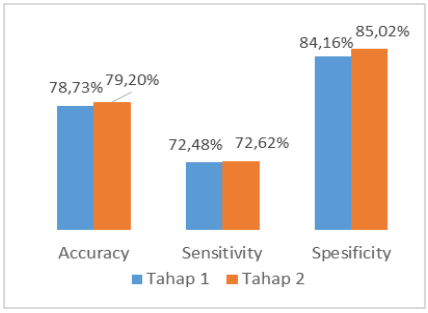

Figure 5. Performance of LASSO logistic regression classification

Variables that Affect on Debtor Status

The variables affecting the debtor's status were analyzed using LASSO logistic regression. The data used in the analysis is all the initial data that has been weighted or WoE calculation. Selection of variables with LASSO logistic regression analysis is done by inserting all explanatory variables and combine two interactions on each explanatory variable. The analysis results of LASSO logistic regression model is as follows:

$$g(x) = ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = 0.296 + 0{,}007 * X_1 + 1.240\,X_3 + 0.693\,X_4 + 0.586\,X_5 + 0.411\,X_6 + 0.473X_{11} + 0.218\,X_{12} + 0.281\,X_{13} + 0.511\,X_{14} + 0.480\,X_{16} + 0.002\,X_1X_3 - 0.019\,X_1X_4 - 7.165\,X_2X_3 - 0.962\,X_3X_5 - 0.845\,X_3X_{14} - 0.017\,X_4X_5 - 0.049\,X_4X_6 - 0.020\,X_4X_9 - 0.104\,X_4X_{13} + 0.411\,X_{14}X_{15}$$

Based on LASSO logistic regression model shows that the variables affecting the debtor's smoothness in meting its obligations are the phone ownership, down payment, loan term, occupation, age, marital status, gender, education level, number of dependents, motorcycle type, interaction between motorcycle type with phone ownership and down payment, interaction between phone ownership with occupation, loan terms, and phone number ownership, interaction between downpayment with loan terms, installments, gender and educational level, ss well as the interaction between occupation with income type**.**

Support Vector Machine (SVM)

In forming the SVM classification model, the explanatory variables used are all of the significant explanatory variables that have been selected by using the LASSO logistic regression model. The results of the analysis using SVM are presented in Figure 6.

Based on Figure 6, it is found that the classification size of each kernel in both stages showed almost the same results. However, the linear function in the first stage has the best classification performance compared to other kernel functions. This is based on the value of accuracy, specificity and sensitivity. The accuracy value of 79.94% indicates that a predicted bad debtor and the reality is bad debtor with a predictable good debtor and the fact is good debtor, then it can be classified appropriately that equal to 79,94%. Furthermore, the specificity value of 79.11% indicates that a predictable good debtor and the reality is classified appropriately of 79.11%. Then the sensitivity value of 79.80% indicates that a predicted bad debtor and the reality is bad debtor is classified appropriately equal to 79.80%.

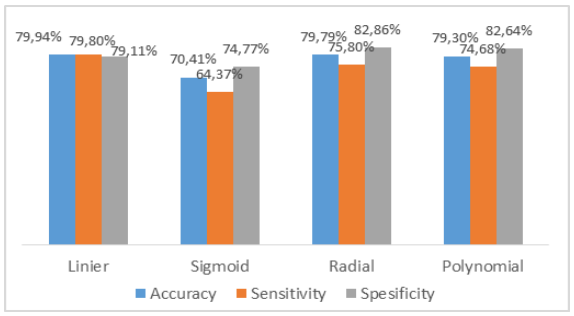

Figure 6. SVM goodness of fit value for each kernel function

Assessment of Logistic Regression Results of LASSO and SVM

The result of classification performance using LASSO logistic regression method suggests that classification accuracy of majority class (spesificity) is 85,02%, higher than that of minority class (sensitivity) that is 72,62%. Meanwhile, the result of classification performance using SVM method shows relatively stable result with classification accuracy value of 79.94% and 79,80% for majority (spesificity) and minority (Sensitivity) classees respectively. Furthenmore, for accuracy value of both methods reveals that the SVM method (79,94%) has a higher than that of LASSO logistic regression method (79,20%). Comparison of classification performance for both methods is shown in Table 2.

Table 2. Goodness of fit classification of LASSO logistic regression and SVM

Goodness	LASSO logistic	SVM
Accuracy	79.20%	79.94%
Sensitivity	72.62%	79.80%
Spesificity	85.02%	79.11%

IV. CONCLUSION

Based on the results and discussion of the classification of the debtor's status as well as the analysis of the variables affecting the debtor's smoothness in meeting its obligations, it can be concluded that:

1. The application of LASSO logistic regression and SVM methods in this case has almost the same classification accuracy, yet the result of classification performance using SVM method shows relatively stable result compared to LASSO logistic regression.
2. The result of LASSO logistic regression analysis shows that the variables affecting the debtor's smoothness in meting its obligations are the phone ownership, down payment, loan term, occupation, age, marital status, gender, education level, number of dependents, motorcycle type, interaction between motorcycle type with phone ownership and down payment, interaction between phone ownership with occupation, loan terms, and phone number ownership, interaction between downpayment with loan terms, installments, gender and educational level, ss well as the interaction between occupation with income type**.**

REFERENCES

[1] www.bi.go.id
[2] Osborne, M.R., Presnell, B., dan Turlach, B.A., 2000, On the LASSO and its dual, J. Comput. Graph. Stat., 9, 319–337.
[3] Steinberg D, Colla P. 1995. CART: Tree-structured nonparametric data analysis. San Diego: Salford System
Agresti, A. 2002. Analysis of ordinal categorical data. Canada: Publisher Simultaneousty.
[4] Karatzoglou, A., Mayer, D., dan Hornik, K. 2006. Support vector machines in r. *jurnal of statistical software. Vol. 15 Issue 9.* http://www.jstatsoft.org/v15/i09/paper. Tanggal Akses 10 Maret 2017.