- Word Count: 3406

Plagiarism Percentage

12%

sources:

1 8% match (Internet from 12-Dec-2014)
http://edoc.ub.uni-muenchen.de/17143/1/Kraus_Anne.pdf

2 1% match (Internet from 29-Mar-2015)
http://en.wikipedia.org/wiki/General_linear_model

3 1% match (Internet from 23-Feb-2017)
http://www.sjournals.com/index.php/SJPAS/article/download/2135/1736

4 < 1% match (Internet from 14-Nov-2017)
https://era.library.ualberta.ca/files/qj72p865r/Alipour_Anahita_Fall%202013.pdf

5 < 1% match (Internet from 24-Jan-2007)
http://www.epicweb.org/legvotes.htm

6 < 1% match (Internet from 02-Nov-2017)
http://repository.petra.ac.id/15818/1/Publikasi1_94032_91.pdf

7 < 1% match (publications)
Li, Chuanxiang, Bin Dai, Tao Wu, and Yiming Nie. "Multi-lane detection in urban driving environments employing omni-directional camera", 2014 IEEE International Conference on Information and Automation (ICIA), 2014.

8 < 1% match (publications)
Siana Halim. "Statistical analysis on the intellectual capital statement", Journal of Intellectual Capital, 2010

9 < 1% match (Internet from 30-Jun-2017)
http://cmapspublic2.ihmc.us/rid=1R4MDDG74-KVX6GG-3L2D/2016%20-

10 < 1% match (publications)
Fan, Xiao, and Lukasz Kurgan. "Accurate prediction of disorder in protein chains with a comprehensive and empirically designed consensus", Journal of Biomolecular Structure and Dynamics, 2014.

11 < 1% match (publications)
Lecture Notes in Computer Science, 2016.

paper text:

Credit Scoring Refinement using Optimized Logistic Regression Hendri Sutrisno

Industrial Engineering Department Petra Christian University Surabaya, Indonesia hensoee@yahoo.com **Siana Halim** 8

Industrial Engineering Department Petra Christian University Surabaya, Indonesia halim@petra.ac.id **Abstract—** 6

A poor credit scoring model will give a poor power for predicting defaulted loan. There are many approaches for modeling the default prediction, such as classical logistic regression and Bayesian logistics regression. In this paper, we applied both classical logistic regression and AUC (Area under Curved) optimized using Nelder-Mead Algorithm for refining a credit scoring model that has already been used for several years by an International bank in Indonesia. Both classical logistics regression and AUC optimized method perform well in improving the model, but logistic regression still better in some aspects. AUC Optimized model has higher AUC than logistic regression model but has lower KS-Score Keywords— Credit scoring, logistics regression, Nelder-Mead Algorithm, AUC optimization I. INTRODUCTION In Indonesia, SMEs (Small and Medium Enterprises) constantly contribute more than 57% in Gross Domestic Product since 2006 [1]. Until 2013, there were more than 57 million SMEs in Indonesia [2]. Every year, Bank X receives thousands of SMEs loan applicant and, as a result, it needs a tool that can process the loans faster and provide low risk. Credit scoring helps lenders take faster, cheaper, and more objective decisions in terms of providing loans [3]. Every classification technique

for credit scoring data gives different results, **where** neutral **networks and least-squares support vector machines yield good results, but the classical logistic regression** is **still** performing **well for credit scoring[** 1

4]. Until now,

logistic regression remains **the main method applied in the banking sector to develop the scoring models.** Since **the** 1

market is changing rapidly, new methods are required for optimizing the scoring problem. 1

In recent years, many quantitative techniques have been used to examine predictive power in credit scoring 1

[5]. Credit scoring models are usually evaluated using power curve such as the Receiver Operating Characteristic (ROC) curves [6].

AUC is an **area under the ROC curve** 10

and a good ROC should have high AUC value. Higher AUC mean the model better in predict bad debtors. Both the ROC curve and the AUC do not depend on the proportion of defaulters in the credit portfolio, therefore they could be used to monitor the performance of credit models over time [7]. Kraus [8], tried to optimize AUC it

seems to be a reasonable procedure for estimating the parameters for credit scoring case 1

besides logistic regression. This research will focus on how to validate current credit scoring model of an International Bank in Indonesia. When the model has already been validated, another interesting problem is how to develop a better classifier. So this research also focusses on developing the credit scoring model using AUC Optimization. II. METHODS A. Logistics

Regression Logistic regression is a statistical method for analyzing dataset in which there are one or more independent variables that determine outcome, which **is** 3

only have two outcomes [9].

In retail banking, logistic regression is the most widely used method for classifying applicants into risk classes because of its **good interpretability and simple explanation** 1

[10]. Logistic regression model is built with a modification of linear regression. = 0 + 11 + 22 + + + (1) Equation (1) considers observation

of one dependent variable and independent variables. Thus, is the h observation of the dependent variable, variable **is h observation of the h** 2

independent variable,

start from 1 to

. The values of **βj represent parameters to be estimated. The** 2

value of Yi will be between -∞ and +∞ depends on the value of independent variables. In order to make the value of always positive, the value will only range between 0 and +∞. To transfer the value of Yi into a range between 0 and 1, then the binary logistic regression is used as the transferred function π(x) = exp Yi (2) 1+exp Yi Therefore, the formula of logit transformation would become as below.

g(x) = ln [π(x)] 1− π(x) 4

(3) Using the logit transformation formula, we can turn back the logistic regression to linear regression B. AUC Optimization – Nelder Mead Algorithm Default customers are customers who fail to pay installments for the loan, and Non-Default customers are customers who pays regular installments for the loan. These

classes are used for the description of the ROC graph. 1

If a default is correctly classified and predicted as a default, it is a true positive( ); **while a non-default wrongly predicted as a default is counted as a false positive** 1

().

TP rate = default **correctly classified (tp)** (4) **total** default **(p) FP rate = non−** default **incorrectly classified (fp)** (5) **total non−** 1

default (p) ROC curve is created

by plotting TPR (true positive rate) versus FPR (false positive rate). 7

Figure 1 shows an example of 11

ROC Curve.

AUC is an **area under the ROC curve.** A good **ROC** 4

should have high AUC. Higher AUC mean the model better in predict bad debtors. When the AUC is equal to 1, it becomes the ideal model, which means the model leads to zero FPR or, in other words, there is no non-default debtor that is incorrectly classified. AUC is computed with the following formula [8] [11]: AUC = 1 nnd . nd ∑n1nd ∑n1d S(xnd, xd) (6) Eguchi and Copas [11] started

AUC optimization with **linear scores by dealing with a complex calculating method for the AUC.** 1

Kraus [8] has proposed a recent method for building credit scoring model, which is called AUC optimization, with

Wilcoxon Mann-Whitney procedure as method of calculation and Nelder-Mead method as the optimization algorithm. The outcomes are compared using different performance measures, and DeLong's test for **analyze the significance of the** different **AUC measures.** Extending **the** 1

definition of equation (6);

is introduced as a vector of coefficients, while nnd **and nd denote the scores as vectors of explanatory variables: AUC(β) = 1** 1

∑n1nd ∑n1d S (βt(xnd, xd)) (7) nnd . nd The

aim is to optimize the β-coefficients by maximizing ( ): arg max **β̂ AUC** 1

= 1 β nnd . nd ∑n1nd ∑n1d S (βt(xnd, xd)) (8) III. RESULTS AND DISCUSSIONS A. Model 2.0 Model 2.0 consider 31 predictor variables. The dataset has 14,700 responses which consist of 14,290 non-default and 4,410 default. It will be separated randomly into train and test dataset with 70:30 proportions. Training dataset has 248 defaults and 10,042 non-defaults, while testing dataset has 98 defaults and 4,312 non-defaults. Only good Predictor Variables will be selected for the model. Pearson's Chi-Square Test for Independence is used to filter the good predictor variables. The Alpha of the test is 0.1., with the hypothesis test as the follow: Fig 1. Example of ROC curves Fig 2. Model 2.0 score distribution Fig 3. Receiver operating characteristic curve of Model 2.0 H0: Response Variable is dependent with Predictor Variable. H1: Response Variable is not dependent with Predictor Variable. As the result of the test, there are 11 predictors variable that can predict default variable with error of 10% (Table 1). Not all good predictor variables are significant to the model of the model. Only significant predictor variables will be selected for the model. These selected variables should not have multi- collinearity with the others. The predictor variables should not have VIF value more than 5, which indicates that these variables are not multi-collinear with the others

(Table 2). Table 1. Pearson's Chi-Square Test 9

for Independence in Variables Selection Factors Sig. test Explanation Factors Sig. test Explanation RP1 0.283686 Bad SC5 RP2 0.019249 Good FC1 RF1 0.367032 Bad TI QM1 0.089796 Good T2 QM2 0.554222 Bad TI QM3 0.740063 Bad T3 QM4 0.075396 Good TI QM5 0.067597 Good OC1 QM7 0.683416 Bad CR1 QM8 0.443678 Bad CR2 QM9 0.427229 Bad AA1 YO1 0.665467 Bad AA2 SC1 0.297085 Bad LQ1 SC2 0.017499 Good LR1 SC3 0.898905 Bad PR1 SC4 0.934353 Bad Table 2. Model 2.0 Summary Factors Est. value Std. error Z value 0.715114 0.029199 0.175191 0.216289 0.610069 0.622519 0.844008 0.945753 0.936653 0.048498 0.085696 5.00E-05 0.00125 5.00E-05 0.934503 Pr (>|z|)

Bad Good Bad Bad Bad Bad Bad Bad Bad Good Good Good Good Good Bad 5

VIF (Intercept) -4.93731 QM5 -0.24278 SC2 -1.07421 FC1 0.567436 CR2 0.813105 AA2 -2.21651 LQ1 -0.70768 LR1 6.380752 0.552664 0.149042 0.332101 0.26886 0.436951 0.176919 0.177591 0.271464 -8.93366 -1.62897 -3.2346 2.11053 1.860862 -12.5284 -3.98492 23.50493 4.12E-19 0.103319 0.001218 0.034813 0.062764 5.22E-36 6.75E-05 3.6E-122 0 1.002843 1.006764 1.00573 1.003276 1.040163 1.011919 1.051683 Fig 4. Model 2.0 score threshold Fig 5. Model 2.1 summary Fig 6. Model 2.1 score distribution As seen on Fig 2 there is a significant difference scores between default and non-default applicants. Measuring the stability of a population aims to find out whether the population of testing dataset differs from the population of training dataset. Population Stability Index usually used as the indicator that still the development population perform as well as in the validation population. With a very low index of 0.008674917, there is an insignificant change between train and test dataset judged using Model 2.0 from its scores. The ability of scorecard to separate between default and non- default can be measured by calculating the value of Kolmogorov-Smirnov Score. The KS Score of this model is 0.754. This means that Model 2.0 is good enough to separate the defaults and non-defaults. The model is very good in separation. The good separation of Model 2.0 also can be seen by the AUC (see Fig. 3). Applicants with score lower than or equal to 52.3 will be predicted as default, while those with score higher than 52.3 will be predicted as non-default. The classification of Model 2.0 correctly predicts 84.1% the default and 90.2% the non-default (Fig 4). The next section will discuss how to improve the AUC with Nelder-Mead Algorithm. B. Model 2.1: AUC Optimization The objective of this optimization is to get a better AUC by changing the parameter value of predictor variables (Fig 5). The optimized model called Model 2.1. The score distribution of this model can be seen in Fig 6. In this figure we can figure out that there are some customers even though the score is only 35 but they were not defaulted, Fig 7. Model 2.1 receiver operating characteristic curve Table 3. Model 2.0-Model 2.1 comparison KS Score 0.754 0.745 Population Stability Index 0.008 0.001 Error Type I 1,407 applicants 1,320 applicants Error Type II 55 applicants 55 applicants Fig. 9 Model 2.1 true default predicted population Fig 8. Model 2.0 score threshold and some of them defaulted even though the score is 95. The simulation value, AUC score is 0.891 after going through 227 times the function is called. Convcode is an integer code, which 0 indicates successful convergence. There is no significant difference in interpret Model 2.1 and Model 2.0 since there are no changes in the parameter value from true real positive to true real negative or otherwise (Fig 7). As seen on Fig 6 there is a significant difference scores between default and non-default applicants. With a very low index of 0.000456, there is an insignificant change between train and test dataset judged using Model 2.1 from its scores. With KS Score of 0.745, Model 2.1 is good enough to separate the defaults and non-defaults. The model is very good in separation. The defaults and non-defaults score distribution is quite good in separating the defaults and non-defaults. The good separation of Model 2.1 also can be seen by the AUC. The AUC of Model 2.1 is 0.881. Model AUC is a very good model in separating defaults and non-defaults. There is a small improvement of AUC Score from Model 2.0 by 0.005. The model is very good in separation. Applicants that have score lower than or equal to 52 will be predicted as default while those with score higher than 52 will be predicted as non-default. From all the 1,611 applicants that Fig. 10. Model 2.1 score below threshold groups scored lower than or equal to 52, 291 were default while 1,320 are not. From all 13,089 applicants with score higher than 52, 13,034 were not default while only 50 default (see Fig 8). C. Model 2.0 – Model 2.1 Comparison There are differences between Model 2.0 and Model 2.1. To select the best one, we can compare those Models by some critical aspects. Table 3 shows the comparison. The AUC value of Model 2.0 and Model 2.1 are good, which Model 2.1 has a better score. The KS Score of Model 2.0 and Model 2.1 are same good which Model 2.0 has higher value than Model 2.1. The Population Stability Index of Model 2.0 also higher than Model 2.1 where Model 2.1 almost considered that a little potential that the population of training dataset is different with the testing dataset. Model 2.0 give a higher type 1 error but have a same type 2 error with Model 2.1. Model 2.0 is not a bad model, but Model 2.1 perform better with the higher score of AUC, lower index of population stability and also a lower error in type I error. D. Error Analysis There are two categories of error of Model 2.1. The first is type I error, an error occurs when the prediction is default but actually non-default. This error will give a potential lost for Bank X. Bank had already rejected the applicant because of the poor score, but actually the applicant is not default, so Bank X will suffer loss by this error. The second one is type II error in which the prediction is non-default but actually default. This kind of error can also cause the Bank to suffer, because the Bank had already accepted the applicant loan in prediction the applicant will not become default, but actually the applicant is default. The chance of type II error in Model 2.1 only about 0.42% (50 cases of 13,089 applicants but 81.93% for type I error (1,320 cases of 1,611 applicants). Model 2.1 contains a very high type I error and offers many risk in its application. The risk of type I error is Bank X lose the potential income applicants which not default. Deep analysis is needed to reduce the risk of the implementation of Model 2.1. Some non-default applicants may have a lower attribute score in the selected variables which lead to score that below the threshold. These applicants are a potential applicant that can bring benefit for Bank X. We can know the applicants who truly default by their score attribute population. The population of the true predicted default applicants (applicant who is predicted as default and actually default) as in the Fig 9 The AUC of Model 2.0 is 0.876. Model AUC is a very good model in separating defaults and non-defaults. Using the discriminant analysis, we can know the threshold score for accepting or rejecting applicants (Fig 10). After knowing the distribution of the true predicted default applicants, we can see the applicants in that population and its score. There are 772 of 1,611 applicants that categorized to this group. 30.05% applicants who match with the first group were default. 232 of 291 (79.7%) of true default predicted applicants were in this group. The interesting point is only 7.03% from the second group were default. 92.97% applicants of this group were not default. This system should be make Bank X more easy to take further action for them who have score below the threshold. This system offers lower risk in accepting or rejecting the applicants who have score below the threshold than Model 2.1. CONCLUSION Recently there are many techniques in developing a good credit scoring model, such as AUC Optimization. AUC is a value that indicate how good a model in separating two different populations. Higher AUC value mean the model better in separating the populations. By optimizing the AUC value, there is a possibility that the AUC value get higher. In term of optimizing, there are many techniques in optimizing, such as Nelder-Mead Algorithm. The first used model in term of building a good model is Model 2.0 based on Logistic Regression Model. Model 2.0 performs well and is able to predict the default and the non- default applicants accurately with AUC value of 0.876. The second one is Model 2.1 based on AUC Optimization from Model 2.0. Model 2.1 has a 0.881 AUC value which is higher than Model 2.0's AUC value, but has a lower KS-Score. Although has a lower score, Model 2.1 still remain in the same class with Model 2.0 of their KS Score. Model 2.1 also has a lower type I error. Model 2.0 predict 1,407 applicants as default applicants but they are actually not going to default, while Model 2.1 only 1,320 applicants. Even Model 2.1 performs better; the model still has a very large of errors. Model 2.1 can only predict 18.06% default applicants correctly which is mean that Model 2.1 also contain a lot of risk in its implementation. Model 2.1 rejected about 81.94% of all applicants with score below the threshold while they were not default. Bank X could lose many potential incomes. Deep analysis can provide some options to reduce the risks of the implementation of Model 2.1. By knowing the true default applicants' population, we can filter the default predicted applicants into two groups. The first group is the default predicted applicant who have a similar population to the population of the true default applicants. The second one is another group except the first population in default predicted applicants. With this separation, Model 2.1 has type I error of 69.95% for the first group and 92.97% for the second group. If the applicant has a score below the threshold and included to the second group, then Bank X has a lower risk in accepting the applicants from 18.06% to 7.03%. Bank X also has a lower risk in rejecting the applicants if the applicant matched with the first group. Model 2.1 can only predict the true default applicants with a rate of 18.06 and if we added the first group as additional filtering system, then Model 2.1 can predict the true default applicants with a rate of 30.05% REFERENCES [1] Statistics Indonesia, "Statistics Indonesia". Jakarta: Statistics Indonesia, 2015. [2] State Minister for Cooperatives Small and Medium Enterprises of Indonesia. "Data Usaha Mikro, Kecil, Menengah (UMKM) dan Besar". Jakarta: State Minister for Cooperatives Small and Medium Enterprises of Indonesia, 2013. [3] L, Mester, "What's the Point of Credit Scoring?", Philadelphia: Federal Reserve Bank of Philadelphia, 1997. [4] B. Baesens, T. Gestel, S. Viaene, M. Stepanova, J. Suyken, J., and J. Vanthienen, "BenchmarkingState-of-the-ArtClassificationAlgorithmsforCreditScoring".The JournaloftheOperationalResearchSociety,2003,pp.627-635 [5] Thomas, L., Oliver, R., & Hand, D. (2005). A Survey of the Issues in Consumer. Journal of the Operational Research Society, 1006-1015. [6] W. Kraznowski, and D. Hand," ROC Curves for Continuous Data". Boca Raton, Florida, United States of America: Taylor & Francis, Inc., 2009. [7] J. Sobehart, and S. Keenan, "Measuring Default Accurately". Credit Risk Special Report, 2001, pp. 31-33. [8] A. Kraus, "Recent Methods from Statistics and Machine Learning for Credit Scoring". Munich, Germany: Ludwig Maximilian University of Munich, 2014. [9] D. Hosmer, D. and S. Lemeshow, Applied Logistic Regression. United States of America: John Wiley & Sons, Inc., 2000. [10] D. Hand, and M. Kelly, "Super Scorecards". IMA Journal of Management Mathematics 13, 4,2002, pp. 273-281 [11] S. Eguchi, S. and J.. Copas, "A Class of Logistic-Type Discriminant Functions". Biometrika 89, III,2002, pp. 1-22.