International Journal of Scientific Research and Engineering Development-– Volume 5 Issue 1, Jan-Feb 2022

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Application of Logistic Regression Models in Credit Scoring: A Case Study of Home Equity Loans

Hoang Thanh Hai*, Do Thanh Phuc**

*(Basic Science Department, Thai Nguyen University of Economics and Business Administration, Viet Nam
Email: hoangthanhhai03091988@gmail.com)

** (Basic Science Department, Thai Nguyen University of Economics and BusinessAdministration, Viet Nam
Email: thanhphuc@tueba.edu.vn)

----------**********----------

Abstract:

In this paper, we use logistic regression to construct a customer classification model based on the data of 5960 home equity loans. This model is used to assess the correlation between customers' characteristics and the probability of their loans to be bad. Finally, we examine the benefits of banks when using this model in terms of profit.

Keywords **—credit scoring, logistic regression, home equity loan.**

----------**********----------

I. INTRODUCTION

Credit scoring is an analysis performed by lenders to determine the creditworthiness of an obligor. Based on customers' information, for instance, age, gender, income, employment status, or credit history, banks score credit applications and ultimately decide which ones to accept and which to reject.

There are two main approaches to assessing credit risk: the judgmental approach and the statistical approach. The former is a qualitative approach. The credit expert or credit committee, based on business experience and common sense, will make a decision about the credit risk. The statistical approach uses historical data to find the optimal multivariate relationship between a customer's characteristics and the binary good/bad target variable. It is less subjective than the judgmental approach. It is also better in terms of speed and accuracy. This is especially relevant when working in an online environment where credit decisions need to be made quickly. Because of these advantages, statistical credit scoring models are considered superior and becoming more and more popular.

Numerous statistical methods have been used to support the credit approval decision process. Among them, logistic regression is one of the most commonly used data mining techniques to conduct credit scoring models thanks to its classification capability and its easy-to-use aspect. In the logistic regression model, one assumes that the probability of a bad loan is given by

$$p(\boldsymbol{x}) = P[Y = 1|\boldsymbol{x}] = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}},$$

or equivalently,

$$\ln\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p,$$

where

x_i: relevant characteristic, $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$,

b_i: corresponding weight;

Y: dependent variable, $Y = 1$ if the loan is classified as "BAD", $Y = 0$ if the loan is classified as "GOOD".

The weights b_i are estimated by the maximum likelihood method. A new loan is allocated to the population of the good loans if its predicted probability $p(\boldsymbol{x})$ is higher than a cut-off level c, which will be determined to maximize the model's accuracy and the bank's profit.

The primary aims of this research are:

- Conduct a credit scoring model using logistic regression;
- Assess the relationship between a client's profiles and his/her probability of being classified as bad credit;
- Determine a cut-off level c to maximize the bank's profit.

II. THE METHODOLOGY

2.1 DATA

The study uses data on 5960 home equity loans. A home equity loan is a loan where the obligor uses the equity of his or her home as the underlying collateral. Dependent variable BAD is binary , $BAD = 1$ ($BAD = 0$) if the applicant is classified as a Bad (Good) credit risk. There are 12 explanatory variables given in Table 1. The total sample contains two kinds of loans: bad loans (1189) and good loans (4771).

2.2 EXPLORATORY DATA ANALYSIS AND DATA CLEANING

We use RStudio software version 1.1.463 for exploratory data analysis and data cleaning. There are 2 categorical independent variables, 6 continuous independent variables, and 4 integer predictors.

2.2.1 Missing Values

Because of a relatively large number of missing values in the data set (Figure 1 and Figure 2), we opt for an imputation method instead of removing missing values from the data. We use K nearest neighbors (KNN) algorithm for imputing missing data with $K = 5$. Imputed data is saved for the next steps.

Table 1: Code sheet for predictors in the data

	Description	*Types*	*Codes/Values*
1	Amount of the loan request	numeric	LOAN
2	Amout due on existing mortage	numeric	MORTDUE
3	Value of current property	numeric	VALUE
4	Reason	categorical	REASON DebtCon = debt consolidation HomeImp = home improvement
5	Occupational categories	categorical	JOB Mgr, Office, Other, ProfExe, Sales, Self
6	Years at present job	numeric	YOJ
7	Number of major derogatory reports	integer	DEROG min. : 0, max. : 10
8	Number of delinquent credit lines	integer	DELINQ min. : 0, max: 15
9	Age of oldest credit line	numeric	CLAGE (months)
10	Number of recent credit inquiries	integer	NINQ min. : 0, max.: 17
11	Number of credit lines	integer	CLNO min. : 0, max: 71
12	Debt-to-income ratio	numeric	DEBTINC

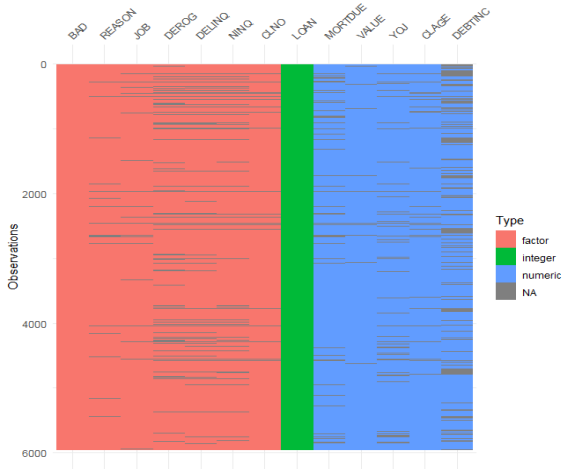

Fig. 1. Missing observations in the data set

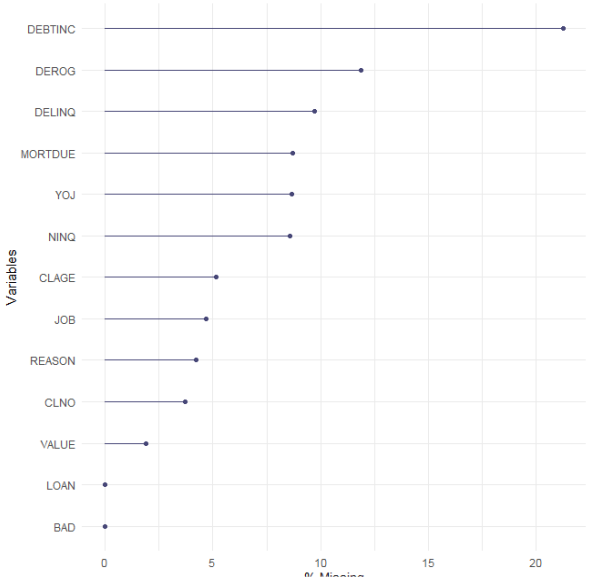

Fig. 2. Percentages of missing values

2.2.2 Categorical variable anaysis

Mosaic plots and Pearson tests are used to analyze categorical independent variables and their influences on the dependent variable. Predictors that have p values in Pearson tests greater than 0.1 will be eliminated when fitting the model. Both two categorical variables have small p-values, so none is deleted. (Figure 3)

Next, we fit a univariable logistic regression model for each categorical covariate. The results of this analysis are shown in Table 2. Note that in this table, each row presents the results for the estimated regression coefficients from a model containing only that covariate.

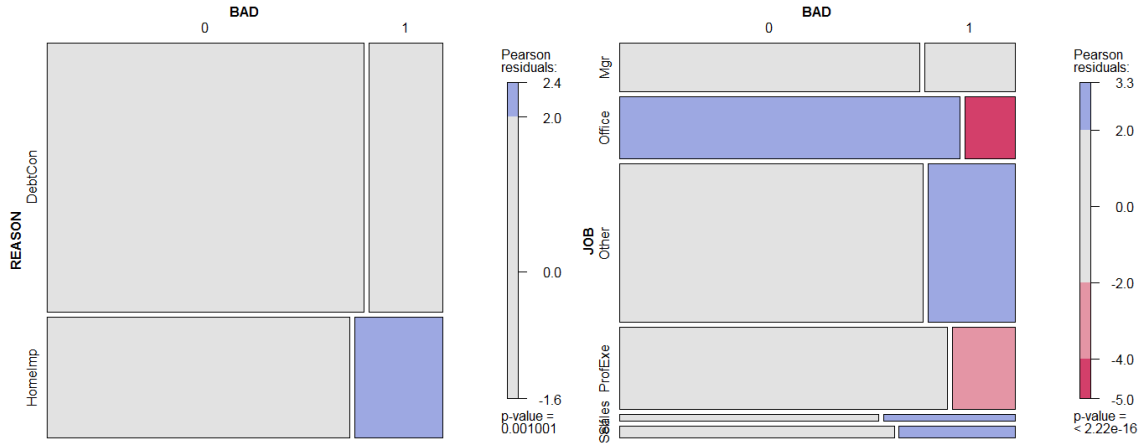

Fig. 3: Mosaic plots of categorical predictors

Table 2: Results of Fitting Univariable Logistic Regression Models
(Categorical Predictors – before Collapsing Categories)

	Coeff.	Std. Err.	OR	Ref.	95% CI	p – value
REASON						
HomImp	0.2253	0.0696	1.25	DebtCon	(1.09, 1.43)	0.001
Job						
Office	-0.7141	0.1268	0.49	Mgr	(0.38, 0.63)	<0.001
Other	-0.0411	0.0968	0.96		(0.80, 1.16)	0.6712
ProfExe	-0.4458	0.1127	0.64		(0.51, 0.80)	<0.001
Sales	0.5224	0.2162	1.69		(1.10, 2.56)	0.016
Self	0.3428	0.1779	1.41		(0.99, 1.99)	0.054

We can see that there is no difference between the probability of a bad loan in group JOB = Other and

in group JOB = Mgr. The coefficient of Other in the model is insignificant. Hence, we merge two levels Other and Mgr into a new level named Other. The result of this combination is shown in Table 3.

Table 3. Results of Fitting Univariable Logistic Regression Model (Predictor JOB – after Collapsing Categories)

	Coeff.	Std. Err.	OR	Ref.	95% CI	p - value
Office	- 0.6829	0.1033	0.51	Other (Mgr – Other)	(041, 0.62)	< 0.001
ProfExe	- 0.4145	0.0855	0.66		(0.56, 0.78)	< 0.001
Sales	0.5537	0.2034	1.74		(1.16, 2.57)	0.007
Self	0.3740	0.1620	1.45		(1.05, 1.99)	0.021

2.2.3 Numeric Variables

a. Outliers

The first step in analyzing continuous predictors is to identify outliers. In general, x is called an outlier of a sample if $x > Q_3 + 1.5.IQR$ or $x < Q_1 - 1{,}5.IQR$, where, Q_3, Q_1 are the third and the first quantiles of the sample, respectively and $IQR = Q_3 - Q_1$.

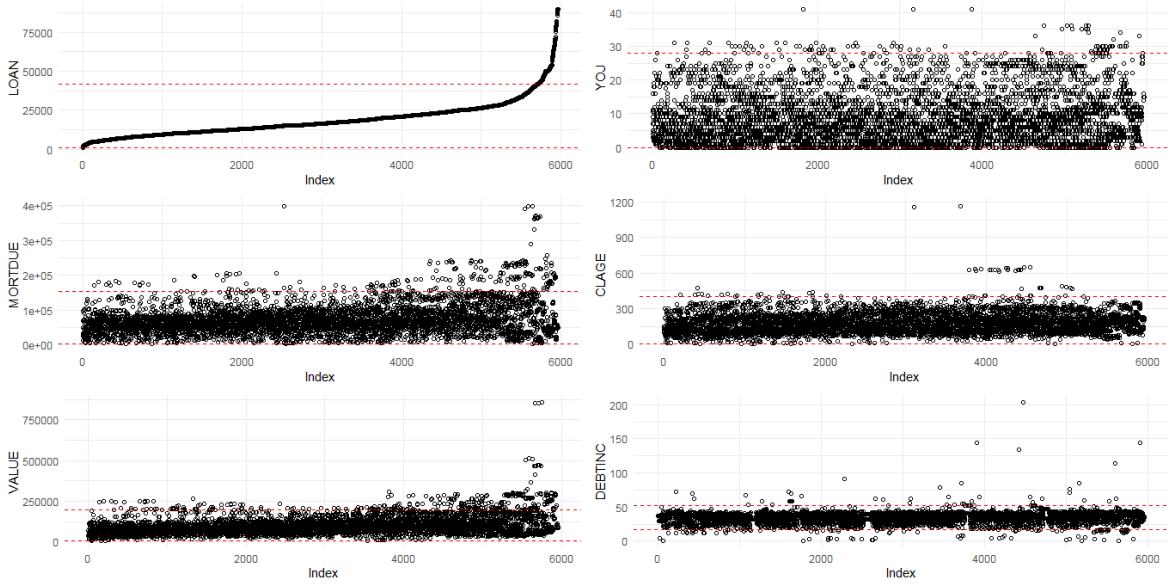

Figure 3: Scatter plots of continuous predictors

Figures 3 shows scatterplots of five continuous variables in the data. The two red dashed lines in each scatterplot represent values $Q_3 + 1.5.IQR$ and $Q_1 - 1{,}5.IQR$. Although there are a few outliers of DEBTINC satisfying the latter condition, the lower bound seems to be too high. So we would not remove these observations. It looks like the upper bound values are too low. Therefore, we increase the upper bounds of LOAN, MORTDUE, VALUE, YOJ, CLAGE and DEBTINC to 75000, 300000, 500000, 35, 600, and 100, respectively. After exploration of three variables for outlierswe have collected 83 indexes to remove. Number of relevant unique observations to remove is 80. Hence, our final data has 5880 observations.

b. Violin Plots and Boxplots

Next, violin plots and fitting univariable logistic regression models are used to assess the impact on the response variable. For example, Figure 4

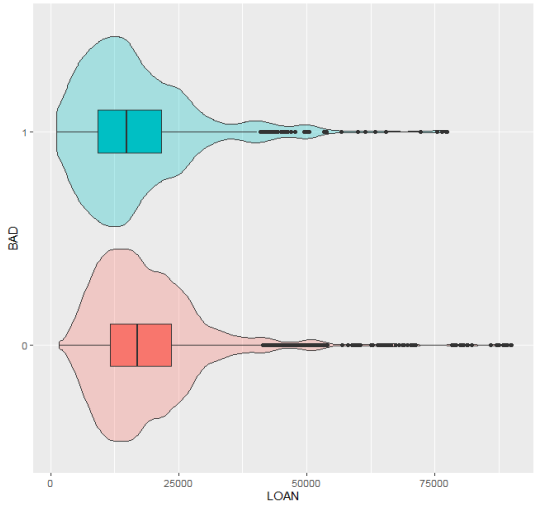

Fig. 4: Violin plot for LOAN

illustrates the impact of LOAN on probability of a bad loan. We could see that the loan in group BAD (BAD =1) tends to lower than that in group GOOD (BAD = 0). In general, the probability of a bad loan decreases when the loan increases. Table 5 shows the result of fitting univariable logistic regression models forall numeric covariates.
CLNO will be removed when fitting model because of its big p – value (0.694).

2.3 Buliding a logistic regression model

After exploratory data analysis and data cleaning process, our final data remains 11 independent variables and 5880 observations. These observations are randomly partitioned into two equal sized subsets – Training (2940) and Testing (2940) data.

The method to be used for the selection of covaiates is Akaike's Information Criterion (AIC). Akaike's Information Criterion of a model is defined as

$$AIC = \frac{-2\ln L + 2(p+1)}{N}$$

Table 5. Results of Fitting Univariable Logistic Regression Models
(Numeric Predictors)

	Coeff.	Std. Err.	OR	Ref.	95% CI	p
LOAN	-0.019	0.003	0.98	1 000 unit increase	(0.97, 0.99)	< 0.001
MORTDUE	-0.037	0.008	0.96	10 000 unit increase	(0.95, 0.98)	< 0.001
VALUE	-0.014	0.006	0.99	10 000 unit increase	(0.97, 1.00)	0.020
YOJ	-0.021	0.005	0.98	1 – year increase	(0.97, 0.99)	< 0.001
CLAGE	-0.078	0.006	0.93	12 – month increase	(0.92, 0.94)	< 0.001
DEBTINC	0.117	0.006	1.12	1 – percent increase	(1.11, 1.14)	< 0.001
DEROG	0.753	0.047	2.12	1 - report increase	(1.94, 2.33)	< 0.001
DELINQ	0.734	0.034	2.08	1 – credit line increase	(1.95, 2.23)	< 0.001
NINQ	0.217	0.017	1.24	1 – inquiry increase	(1.20, 1.28)	< 0.001
CLNO	-0.001	0.003	1.00	1 – credit line increase	(0.99, 1.01)	0.694

where L is the likelihood of the model, p is the number of independent variables in the model and N is the number of obsevations. The model with the smaller AIC is considered the better fitting model. Based on AIC and Training data, we have selected the model with the smallest AIC containing all the above eleven variables: VALUE, REASON, YOJ, MORTDUE, LOAN, NINQ, JOB, DEROG, CLAGE, DEBTINC, DELINQ.

Table 6. Regression results of the final model

	Coeff.	Std. Err.	OR	Ref.	95% CI	p
VALUE	0.049	0.022	1.05	10 000 unit increase	(1.01, 1.10)	0.024
REASONHomeImp	0.309	0.123	1.36	DebtCon	(1.07, 1.73)	0.012
YOJ	-0.028	0.009	0.97	1 - year increase	(0.96, 0.99)	0.002
MORTDUE	-0.080	0.026	0.92	10 000 unit increase	(0.88, 0.97)	0.002
LOAN	-0.023	0.006	0.98	1 000 unit increase	(0.96, 0.99)	< 0.001
NINQ	0.124	0.029	1.13	1 – inquiry increase	(1.07, 1.20)	< 0.001
JOBOther	0.613	0.174	1.85		(1.32, 2.62)	< 0.001
JOBProfExe	0.666	0.205	1.95	Office	(1.31, 2.92)	0.001
JOBSales	1.686	0.403	5.40		(2.42, 11.8)	< 0.001
JOBSelf	1.481	0.323	4.40		(2.32, 8.25)	< 0.001
DEROG	0.509	0.069	1.66	1 - report increase	(1.46, 1.91)	< 0.001
CLAGE	-0.092	0.010	0.91	12 – month increase	(0.89, 0.93)	< 0.001
DEBTINC	0.117	0.010	1.12	1 – percent increase	(1.10, 1.15)	< 0.001
DELINQ	0.749	0.058	2.11	1 – credit line increase	(1.89, 2.37)	< 0.001

III. FINDINGS

3.1 The model

Table 6 shows the results of fitting the final model consisting of eleven above predictors. The values of all coefficients correspond to the exploratory data analysis.

3.2 Assess Discrimination

The classification power of the model is computed on Testing data and is shown in table 7 for a cut-off level c equal to 0.5. The accuracy is 0.86.

Table 7. Classification results of the final model

	Actual	
Prediction	Bad	Good
Bad	243	72
Good	340	2285

Figure 5 shows the performance of the classifier through ROC curve. The area under the curve is 0.84.

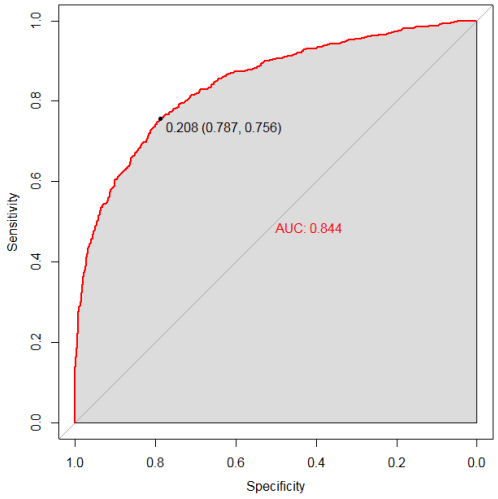

Fig. 5: ROC curve of the final model

3.2 Profit consideration

Let us assume that a correct decision of the bank would result in 30% profit. A correct decision here means that the bank predicts an application to be good and it actually turns out to be credit worthy.

When the opposite is true, i.e. bank predicts the applicant to be good but he/she turns out to be a bad credit, then the loss is 100%. If the bank predicts an application to be non-creditworthy, then loan facility is not extended to that applicant and bank does not incur any loss. The cost matrix, therefore, is as shown in Table 7.

Table 7. Cost matrix

	Actual	
Prediction	Bad	Good
Bad	0	0
Good	-1	0.3

Out of 5960 home equity loans, 4771 loans are good. A loan manager without any model would incur

$$\frac{4771}{5960}.0{,}3+\frac{1189}{5960}.(-1)=0.04$$

unit profit. If the bank uses this model, its per applicant profit would be

$$\frac{2285}{2625}.0{,}3+\frac{340}{2625}(-1)=0.13.$$

Figure 6 shows the profits of the model corresponding to different cutoff levels. Among the five thresholds, the maximum profit is 0.13 per applicant at $c=0.5$.

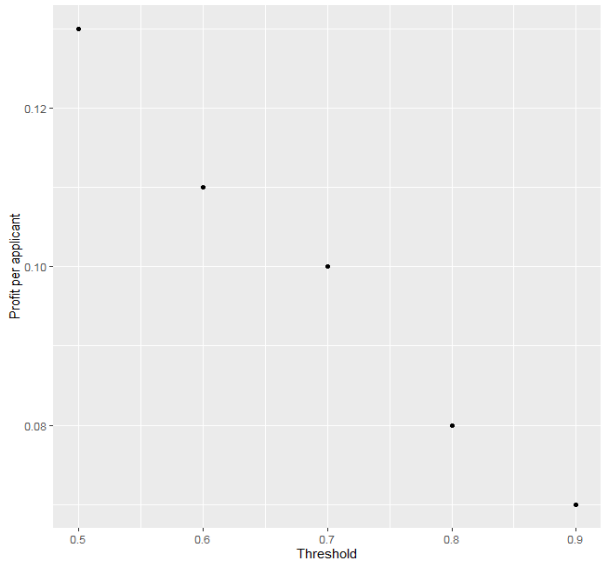

Figure 6: Bank's profits corresponding to different thresholds

IV. CONCLUSIONS

In this paper, we use a logistic regression model to build a credit classification model. The mode's accuracy is 0.7680 and the area under the curve is 0.7753. Also, the predictors' influences on the response variable are analyzed through exploratory data analysis and are confirmed by the results of fitting the final model. Finally, we assess the bank's profit when applying the model corresponding to different thresholds.

REFERENCES

[1] A Steenackers MG, *A credit scoring model for personal loans*, Insurance: Mathematics and Economics, 1989, vol. 8, pp. 31 -34.

[2] Bart Baesens, HS Daniel Rösch, *Credit risk analytics: measurement techniques, applicationsm, and examples in SAS*, John Wiley & Sons, Hoboken, New Jersy, 2016.

[3] Long JS, *Regression models for categorical and limited dependent variables*, SAGE Publications, Thosand Oaks, California, 1997.

