Application Of Logistic Regression Model In Consumer Loans Credit Scoring

Hoang Thanh Hai, Dong Thi Hong Ngoc

Thai Nguyen University of Economics and Business Administration,
Thai Nguyen, Vietnam,
hoangthanhhai03091988@gmail.com

Thai Nguyen University of Economics and Business Administration,
Thai Nguyen, Vietnam,
dongngoc.1088@gmail.com

Abstract: Credit scoring is one of the most crucial processes in banks' credit management decisions. Various scoring techniques have been suggested to assess clients' creditworthiness during the last few decades. In this paper, we use logistic regression to construct a classification model based on data on 1000 loan applicants in Germany. This model is used to examine the correlation between customers' characteristics and the probability of their loans to be good. Finally, we assess the benefits of banks when using this model in terms of profit.

keywords: credit scoring model, logistic regression, probability of a loan to be good, profit

1. INTRODUCTION

The objective of credit scoring models is to assign credit applicants to either a "good credit" group who are likely to repay the financial obligation, or a "bad credit" group who are more likely to default on the financial obligation. This classification helps a bank make decisions regarding who to give approval of the loan and who not to. Two types of risks are associated with the bank's decisions. If the applicant is a good credit risk, then not approving the loan to the person results in a loss of business to the bank; if the applicant is a bad credit risk, then approving the loan to the person results in a financial loss to the bank. Using a credit scoring model with high classification accuracy could reduce credit risk and increase profit for banks. Numerous statistical methods and artificial intelligence approaches have been proposed 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 good 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 "Good", $Y = 0$ if the loan is classified as "Bad".

The weights b_i are estimated by use of 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 as well as 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 demographic and socio-economic profiles and his/her probability of being classified as good credit;
- Determine a cut-off level c to maximize the model's accuracy as well as the bank's profit.

2. THE METHODOLOGY

2.1 Data

The study uses data on 1000 personal loans in Germany. Dependent variable Y is binary, $Y = 1$ $(Y = 0)$ if the applicant is classified as a Good (Bad) credit risk. There are 20 explanatory variables given in Table 1. The total sample contains two kinds of loans: good loans (700) and bad loans (300).

2.2 Exploratory data analysis and data cleaning

We use R software version 3.6.3 for exploratory data analysis and data cleaning. There are 17 categorical independent variables and 3 continuous independent variables.

2.2.1 Categorical variable anaysis

Mosaic plots and Pearson tests are used to analyse categorical predictors and their influences on the response variable. Predictors that have p values in Pearson tests greater than 0.1 will be eliminated when fitting the model. We have determined six variables that will be removed from the model including variables 8, 11, 16, 17, 18, 19 in table 1. Table 2 describes p values of these predictors.

***Table 1**: Code sheet for predictors in the data*

	Description	*Codes/Values*
1	Status of existing checking account	A11 = account < 0 DM
		A12 = 0 ≤ account < 200 DM
		A13 = account ≥200 DM or salary assignments for at least 1 year
		A14 = no checking account
2	Duration	Months
3	Credit History	A30 = no credits taken/all credits paid back duly
		A31 = all credits at this bank paid back duly
		A32 = existing credits paid back duly till now
		A33 = delay in paying off in the past
		A34 = critical account/ other credits existing (not at this bank)
4	Purpose	A40 = car (new)
		A41 = car (used)
		A42 = furniture/equipment
		A43 = radio/television
		A44 = domestic appliances
		A45 = repairs
		A46 = education
		A47 = vacation
		A48 = retraining
		A49 = business
		A410 = others
5	Credit Amount	DM
6	Savings account/bonds	A61 = amount < 100 DM
		A62 = 100 ≤ amount < 500 DM
		A63 = 500 ≤ amount< 1000 DM
		A64 = amount ≥ 1000 DM
		A65 = unknown/ no savings account
7	Length of Current Employment	A71 = unemployed
		A72 = < 1 year
		A73 =1 ≤ length < 4 years
		A74 = 4 ≤ length < 7 years
		A75 = length ≥ 7 years
8	Installment rate in percentage of disposable income	1, 2, 3, 4
9	Personal status and sex	A91 = male: divorced/separated
		A92 = female:divorced / separated / married
		A93 = male: single
		A94 = male: married/widowed
		A95 = female: single
10	Other debtors / guarantors	A101 = none
		A102 = co-applicant
		A103 = guarantor
11	Length of present residence	1,2,3,4
12	Property	A121 = real estate
		A122 = if not A121 : building society savings agreement/life insurance
		A123 = if not A121/A122 : car or other, not in predictor 6
		A124 = unknown / no property
13	Age	Years
14	Other installment plans	A141 = bank
		A142 = stores
		A143 = none
15	Housing	A151 : rent
		A152 : own
		A153 = for free
16	Number of existing credits at this bank	1,2,3,4
17	Job	A171 = unemployed/ unskilled - non-resident
		A172 = unskilled – resident
		A173 = skilled employee / official
		A174 = management/ self-employed/ highly qualified employee/ officer
18	Number of people being liable to provide maintenance for	1, 2
19	Telephone	A191 = none
		A192 = yes
20	Foreign worker	A201 = yes
		A202 = no

***Table 2.** Categorical variables not included in the model (p –value > 0.1)*

	Description	*p - value*
1	Installment rate in percentage of disposable income	0.14
2	Length of present residence	0.86
3	Number of existing credits at this bank	0.45
4	Job	0.60
5	Number of people being liable to provide maintenance for	0.92
6	Telephone	0.25

Next, we fit a univariable logistic regression model for each remaining categorical covariate. The results of this analysis are shown in Table 3. Note that in this table, each row presents the results for the estimated regression coefficients from a model containing only that covariate. In table 3, levels of covariates that have insignificant coefficients will be combined into new categories such that when fitting univariable logistic regression models their coefficients become significant. We use mosaic plots to support this process. For example, Figure 1 shows the mosaic plot of predictor *Length of Current Employment*.

***Figure 1:** Mosaic Plot of Length of Current Employment (before combining levels)*

Table 3. *Results of Fitting Univariable Logistic Regression Models (Categorical Predictors – before Collapsing Categories)*

	Coeff.	Std. Err.	OR	Ref.	95% CI	p – value
Checking Account						
A12	0.4167	0.1739	1.52		(1.08, 2.13)	0.017
A13	1.2236	0.3262	3.4	A11	(1.79, 6.44)	< 0.001
A14	1.9944	0.1980	7.35		(4.98, 10.83)	< 0.001
Credit History						
A31	0.2231	0.4359	1.25		(0.53, 2.94)	0.609
A32	1.2698	0.3396	3.56	A30	(1.83, 6.93)	< 0.001
A33	1.2730	0.3988	3.57		(1.63, 7.80)	0.001
A34	2.0919	0.3616	8.10		(3.99, 16.46)	< 0.001
Purpose						
A41	1.13304	0.29763	3.11		(1.77, 5.72)	< 0.001
A42	0.26364	0.20858	1.30		(0.87, 1.96)	0.206
A43	0.76926	0.19710	2.16		(1.47, 3.19)	< 0.001
A44	0.20505	0.62700	1.23		(0.38, 4.71)	0.744
A45	0.07152	0.46321	1.07	A40	(0.44, 2.78)	0.877
A46	-0.24694	0.31512	0.78		(0.42, 1.46)	0.433
A48	1.59134	1.06915	4.91		(0.88, 91.86)	0.137
A49	0.12868	0.25183	1.14		(0.70, 1.88)	0.609
A410	-0.15163	0.60082	0.86		(0.27, 3.00)	0.801
Savings						
A62	0.13181	0.22606	1.14		(0.73, 1.78)	0.560
A63	0.97741	0.34255	2.66	A61	(1.36, 5.2)	0.004
A64	1.36997	0.44461	3.94		(1.65, 9.41)	0.002
A65	0.97560	0.21230	2.65		(1.75, 4.02)	< 0.001
Length of Present Employment						
A72	-0.1516	0.3053	0.86		(0.47, 1.56)	0.620
A73	0.2871	0.2881	1.33	A71	(0.76, 2.34)	0.319
A74	0.7136	0.3196	2.04		(1.09, 3.82)	0.026
A75	0.5548	0.3001	1.74		(0.97, 3.14)	0.064
Personal Status and Sex						
A92	0.2065	0.3122	1.23	A91	(0.67, 2.27)	0.508
A93	0.6074	0.3044	1.84		(1.01, 3.33)	0.046
A94	0.5804	0.3718	1.79		(0.86, 3.7)	0.119
Other Debtors						
A102	-0.60270	0.32293	0.55	A101	(0.29, 1.03)	0.062
A103	0.58726	0.35925	1.8		(0.89, 3.64)	0.102
Property						
A122	-0.4896	0.2036	0.61		(0.41, 0.91)	0.016
A123	-0.4952	0.1879	0.61	A121	(0.42, 0.88)	0.008
A124	-1.0471	0.2182	0.35		(0.23, 0.54)	< 0.001
Other Installment Plans						
A142	0.0241	0.3436	1.02	A141	(0.52, 2.01)	0.944
A143	0.6048	0.1895	1.83		(1.26, 2.65)	0.001
Housing						
A152	0.59860	0.17531	1.82	A151	(1.29, 2.57)	< 0.001
A153	-0.06816	0.24862	0.93		(0.57, 1.52)	0.784

From Figure 1 we could see that probabilities of a good loan in three groups A71, A72, A73 are roughly equivalent. A similar thing is seen in two groups A74, A75. Overall, the probability of being classified as a good credit risk in customers whose length of current jobs equal to or greater than 4 years exceed that in the remaining customer group. Hence, we combine three levels A71, A72, A73 and two levels A74, A75 into two new categories called Length.current.job1 and Length.current.job2, respectively. Figure 2 shows the mosaic plot of predictor *Length of Current Employment* after merging levels. By this way, we combine levels of other categorical variables. The results of this combination process are shown in Table 4.

Figure 2: *Mosaic Plot of Length of Current Employment (after combining levels)*

Table 4*. Results of Fitting Univariable Logistic Regression Models (Categorical Predictors – after Collapsing Categories)*

	Coeff.	Std. Err.	OR	Ref. (Combined from)	95% CI	p	Combined from
Credit.History2	1.1458	0.2352	3.14	Credit.History1	(1.98, 4.99)	< 0.001	A32
Credit.History3	1.7438	0.2505	5.72	(A30, A31)	(3.5, 9.34)	< 0.001	A33, A34
New.Car	- 0.5235	0.1687	0.59	Home.Related (A42 – A45)	(0.43, 0.82)	0.002	A40
Others	- 0.4753	0.1895	0.62		(0.43, 0.90)	0.012	A46 – A410
Used.Car	0.6095	0.2842	1.84		(1.05, 3.21)	0.032	A41
Savings2	0.95850	0.34106	2.61	Savings1	(1.34, 5.09)	0.005	A63
Savings3	1.35107	0.44346	3.86	(A61, A62)	(1.62, 9.21)	0.002	A64

Savings4	0.95670	0.20989	2.6		(1.73, 3.93)	< 0.001	A65
Length.current.job2	0.49963	0.14329	1.65	Length.current.job1 (A71 – A73)	(1.24, 2.18)	< 0.001	A74, A75
Sex.Marital.Status 2	0.4263	0.1416	1.53	Sex.Marital.Status 1 (A91, A92)	(1.16, 2.02)	0.003	A93, A94
Concurrent.Credit. Yes	-0.59873	0.16855	0.55	Concurrent.Credit. None (A143)	(0.39, 0.76)	< 0.001	A141, A142
Rent.ForFree	-0.62436	0.14774	0.54	Own (A152)	(0.4, 0.72)	< 0.001	A151, A153

After exploratory data analysis using mosaic plots and fitting univariable logistic regression models we have a picture about categorical predictors' impact to response variable as below.

- Checking Account: the probability of a good loan increases significantly from group A11 to group A14.
- Credit History: the probability of a good loan increases significantly from group Credit.History1 to group Credit.History3.
- Purpose: the probability of a good loan in groups New.Car and Others less than that in group Home.Related, the probability of a good loan in group Used.Car greater than that in group Home.Related.
- Savings: the probability of a good loan in all groups Saving2, Saving3, Saving4 is greater than that in group reference Saving1 and the biggest number is in group Saving3.
- Length of current job: the probability of being classified as a good credit risk in customers whose length of current jobs equal to or greater than 4 years exceed that in the remaining customer group.
- Sex Marital Status: the probability of a good loan in group Sex.Marital.Status2 greater than that in group Sex.Marital.Status1
- Other Debtors: the probability of a good loan in group A101 greater than that in group A102 but less than that in group A103.
- Property: the probabilities of a good loan in groups A122 and A123 are almost equivalent, the probability of a good loan in group A121 is smallest.
- Concurrent Credit: the probability of a good loan in the group owning concurrent credit less than that in the remaining group.
- Housing: : the probability of a good loan in the group having own houses greater than that in the remaining group.

2.2.2 Continuous variable anaysis

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 $\mathrm{IQR} = \mathrm{Q}_3 - \mathrm{Q}_1$. Figures 3 shows scatterplots of three continuous variables in the data. The red dashed lines in three scatterplots represent values $\mathrm{Q}_3 + 1.5.\mathrm{IQR}$ (There are no outliers in the data satisfying the latter condition). It looks like the first two values are too low. Therefore, we opt $\mathrm{Q}_3 + 1.5.\mathrm{IQR} + 10$ and $\mathrm{Q}_3 + 1.5.\mathrm{IQR} + 5000$ to be upper bounds for Duration and Credit Amount, respectively. After exploration of three variables for outliers we have collected 53 indexes to remove. Number of relevant unique observations to remove is 47. Hence, our final data has 953 observations. Next, violin plots and fitting univariable logistic regression models are used to assess the influences on the response variable. For example, Figure 4 illustrates the impact of Duration on creditability. We could see that the duration in group Good tends to lower than that in group Bad. In general, the probability of a good loan decreases when duration increases. Table 5 shows the results of fitting univariable logistic regression models for three continuous covariates.

2.3 Buliding a logistic regression model

After exploratory data analysis and data cleaning process, our final data remains 14 independent variables and 953 observations. These observations are randomly partitioned into two equal sized subsets – Training (477) and Testing (476) 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

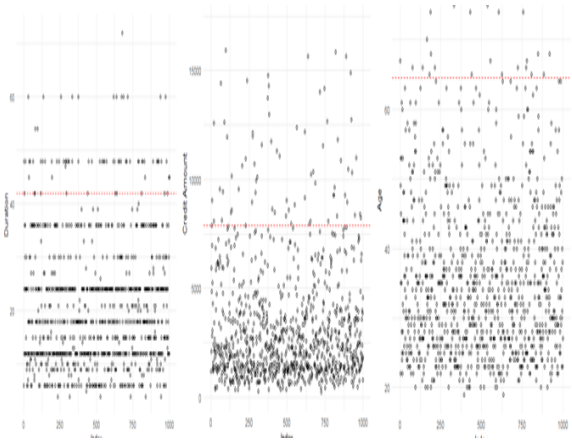

Figure 3: *Scatter plots of three continuous predictors*

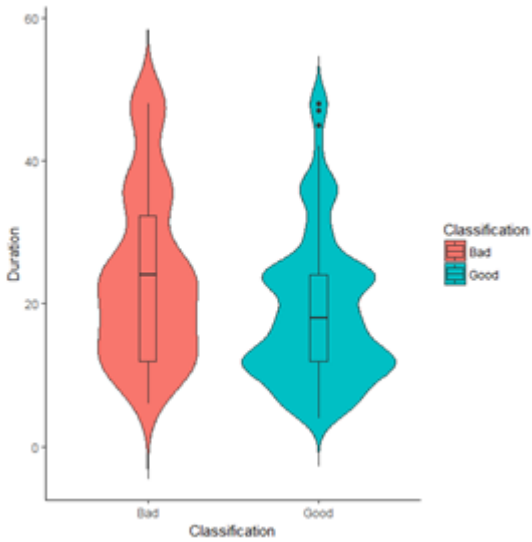

Figure 4: *Violin plot for Duration*

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

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 ten following variables: Status of existing checking account, Credit history, Purpose, Savings, Length of current job, Sex marital status, Other debtors, Concurrent credit, Housing, Duration, Age.

3. FINDINGS

3.1 The model

Table 6 shows the results of fitting the final model including ten aforementioned predictors. The values of all coefficients correspond to the exploratory data analysis. 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 0.7689.

Table 7. *Classification results of the final model*

	Actual	
Prediction	Bad	Good
Bad	62	31
Good	79	304

Figure 5 shows the performance of the classifier through ROC curve. The area under the curve is 0.7753.

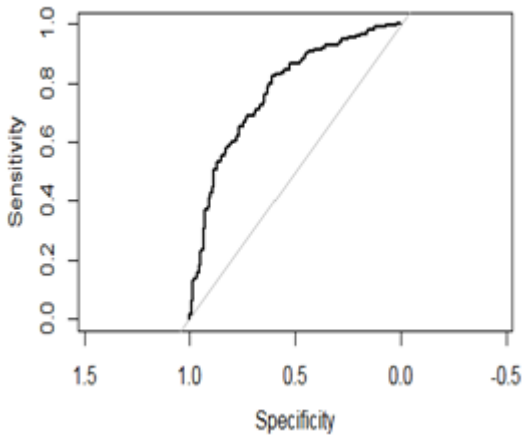

Figure 5: *ROC curve of the final model*

3.2 Profit consideration

Let us assume that a correct decision of the bank would result in 35% profit at the end of 5 years. A correct decision here means that the bank predicts an application to be good or credit-worthy 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 (opportunity loss is not considered here). The cost matrix, therefore, is as shown in Table 7.

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

	Coeff.	Std. Err.	OR	Ref.	95% CI	p
Duration	-0.20387	0.03226	0.82	5 - month increase	(0.77, 0.87)	<0.001
Credit Amount	-0.10032	0.02801	0.90	1000 DM increase	(0.86, 0.96)	<0.001
Age	0.12179	0.03714	1.13	5 - year increase	(1.05, 1.21)	0.001

Table 7. *Cost matrix*

	Actual	
Prediction	Bad	Good
Bad	0	0
Good	-1	0.35

Out of 1000 applicants, 70% are creditworthy. A loan manager without any model would incur $0{,}7.0{,}35 + 0.3.(-1) = -0.055$ or 0.055 unit loss. If the bank uses this model, its per applicant profit would be

$$\frac{304}{476}.0{,}35 + \frac{79}{476}(-1) = 0.058.$$

Figure 6 and Figure 7 show the profits of the bank and the accuracy of the model corresponding to different cutoff levels. Among the five thresholds, the maximum profit is 0.087 per applicant at $c = 0.8$. Meanwhile, at $c = 0.5$ the model' s accuracy would reach the maximum value of 0.7689.

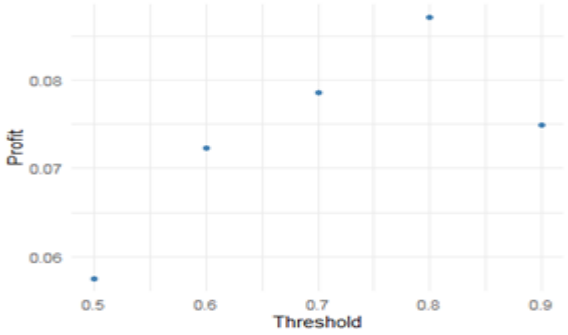

Figure 6: *Bank's profits corresponding to different thresholds*

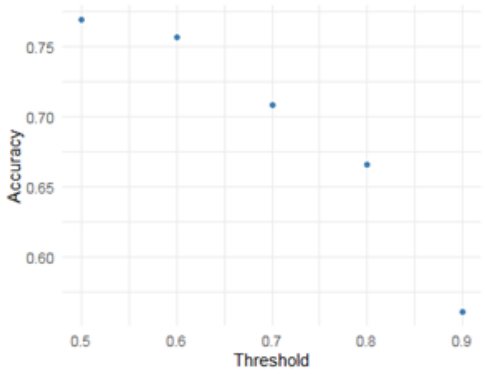

Figure 7: *The model's accuracy corresponding to different thresholds*

4. Conclusions and areas of future research

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.

Table 6. *Regression results of the final model*

	Coeff.	Std. Err.	OR	Ref.	95% CI	p
Checking.AccountA12	0.50655	0.29477	1.66		(0.93, 2.97)	0.0857
Checking.AccountA13	1.18679	0.49281	3.27	Checking.AccountA11	(1.30, 9.18)	0.0160
Checking.AccountA14	1.70442	0.31586	5.50		(3.00, 10.40)	<0.001
Credit.HistoryCredit.History2	0.83411	0.42282	2.30	Credit.HistoryCredit.History1	(1.02, 5.38)	0.0485
Credit.HistoryCredit.History3	1.07507	0.43706	2.93		(1.26, 7.03)	0.0139
PurposeNew.Car	-0.72136	0.29515	0.47		(0.27, 0.87)	0.0145
PurposeOthers	0.23422	0.36591	1.26	PurposeHome.Related	(0.62, 2.63)	0.5221
PurposeUsed.Car	1.00782	0.51618	2.74		(1.04 ,7.95)	0.0509
Savings2	1.61440	0.81191	5.02		(1.25, 34.67)	0.0468
Savings3	1.75940	0.77912	5.81	Savings1	(1.55, 38.21)	0.0239
Savings4	0.43261	0.34211	1.54		(0.80, 3.07)	0.2060
Sex.Marital.Status2	0.54163	0.25303	1.72	Sex.Marital.Status1	(1.05,2.83)	0.0323
GuarantorA102	0.67586	0.70428	1.96	GuarantorA101	(0.53, 8.78)	0.3372
GuarantorA103	0.97054	0.52927	2.64		(0.99, 8.08)	0.0667
Concurrent.Credit.Yes	-0.61332	0.30214	0.54	Concurrent.Credit. None	(0.30, 0.98)	0.0424
HousingRent.ForFree	-0.48233	0.26035	0.62	HousingOwn	(0.37, 1.03)	0.0639
Duration	-0.27013	0.05748	0.76	5-month increase	(0.68, 0.85)	<0.001
Age	0.10931	0.06408	1.12	5-year increase	(0.96, 1.27)	0.0880

REFERENCES

[1]. Harrell, F. E. , *Regression modeling strategies with applications to linear models, logistic and ordinal regression, and survival analysis*, 2nd edition, Springer – Verlag, Cham, 2015.

[2]. Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013), *Applied logistic regression*, 3rd edition, Jonh Wiley & Sons.

[3]. Long, J. S., *Regression models for categorical and limited dependent variables*, Sage Publications, 1997.

[4]. Steenackers, A., Goovaerts, M. J. , *A credit scoring model for personal loans*, Insurance: Mathematics and Economics 8, pp. 31 – 34, 1989.

[5]. Yeh, I. -C., Lien, C. –h. , *The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients*, Expert systems with applications 36, pp. 2473 – 2480, 2009.

[6]. Department of Statistics, Eberly College of Science, Analysis of German Credit Data, *Analysis of German Credit Data*, **https://online.stat.psu.edu/stat508/resource/analysis/gcd**

[7]. Drugov, V. G., *Default payments of credit card clients in Taiwan from 2005*, https://rstudio-pubs-static.s3.amazonaws.com/281390_8a4ea1f1d23043479814ec4a38dbbfd9.html