UNIVERSITY OF TARTU
FACULTY OF SCIENCE AND TECHNOLOGY
Institute of Mathematics and Statistics

Nare Torosyan

Application of binary logistic regression in credit scoring

Financial Mathematics
Master's Thesis (15 ECTS)

Supervisor: Prof. Kalev Pärna

Tartu 2017

Application of Binary Logistic regression in Credit Scoring

Master's Thesis

Nare Torosyan

Abstract. Nowadays, demand for loan products is growing day by day. Also, loan applicants have become more demanding, than they were before. They want to receive the response from bank as soon as possible. In order to resist the growing competition banks develop new quantification techniques which accelerate and automate the decision making process. One of these techniques is credit scoring. Credit Scoring is one of the most widely used instruments which is applied by lenders decide whether to approve or reject the loan application.

In this Master's thesis an overview of credit scoring is given. The most essential objective of this thesis is to show the application of logistic regression in Credit score models. Other methods of credit scoring will also be noted, but not in extensive detail. The study ends with a practical application of logistic regression for a credit scoring model on real data of loan applicants.

CERCS research specialisation: P160 Statistics, operations research, programming, actuarial mathematics.

Keywords: credit scoring, logistic regression.

Binaarse logistilise regressiooni rakendamine krediidiskooringus.

Magistritöö finantsmatemaatika erialal

Nare Torosyan

Lühikokkuvõte. Tänapäeval kasvab nõudlus laenutoodete järele pidevalt. Laenutaotlejad on muutunud nõudlikumaks kui varem. Nada soovivad saada vastust laenutaotlusele nii kiiresti kui võimalik. Et vastu pidada kasvavas konkurentsis, peavad pangad välja arendama uusi kvantitatiivseid tehnikaid, mis kiirendavad ja automatiseerivad laenuotsuste vastuvõtmist. Üheks taoliseks meetodiks on krediidiskooring. Krediidiskooring on üks kõige levinumaid instrumente, mida rakendatakse laenuandjate poolt otsustamaks, kas laenutaotlus rahuldada või lükata tagasi.

Antud töö eesmärk on demonstreerida logistilise regressiooni kasutamist krediidiskooringu mudeli väljatöötamisel reaalsete laenutaotlusandmete põhjal. Esmalt antakse üldine ülevaade krediidiskooringust, seejärel kirjeldatakse logistilise regressiooni mudelit. Logistilise regressiooni mudelit on kasutatud "hea" kliendi tõenäosuse prognoosimiseks. Põgusalt peatutakse ka teistel krediidiskooringu meetoditel.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants-ja kindlustusmatemaatika.

Võtmesõnad: krediidiskooring, logistiline regressioon

Table of contents

1. Introduction

Problem statement: Commercial banks provide financial products and services to their customers. This activity is always accompanied with different types of risk. Risk taking is the essential requirement for the commercial bank's profitability. In other words, current risks may transform to tomorrow's reality. Therefore banks always develop new tools in order to manage all risks effectively.

One of the most common risks among the different banking risks is credit risk. By being exposed to credit risk, banks realized that they need to take conscious risk-taking decisions, which in turn call for quantitative risk management systems. These systems give the bank early warnings for predicting potential failures. The customer's credit risk level is often evaluated by the bank's internal credit scoring models. The aim of these models is to make sophisticated decisions and determine if the applicant has the capacity to do the required payments of the loan. This is usually done by logistic regression using historical data and statistical techniques.

Purpose: The aim of this Master's thesis is to show the role of credit scoring in decision making process, and to demonstrate how the logistic regression works by comparing predicted results with actual results, in order to see how correct the decision rule is.

Definitions of some terms

a) Default

Default is the failure to pay interest or principal on a loan when due.

b) Credit report

Credit report is a detailed report of an individual's credit history, which is collected and created by credit bureaus. Lenders combine this information with other information in order to determine the applicant's credit worthiness.

c) Credit bureau

The credit bureau is an agency that collects and maintains individual credit information, which sells it to lenders so they can make a decision.

d) Score Factors

Score factors are the elements from the credit report that make up the customer's credit score. Examples of such elements are age, gender, marital status, income, late payments,

etc. Score factors can have both positive and negative impact on the person's credit score.

Research method: Mostly quantitative analysis has been used in this Master's thesis. Particularly, binary logistic regression has been implemented on real data of loan applicants.

The Structure of the Thesis: Chapter 2 answers the question what is credit scoring, which are the main types and methods of credit scoring. Chapter 3 gives an overview of logistic regression and describes the evaluation process of the model. Chapter 4 shows how binary logistic regression has been conducted on real data of loan applicants.

2.Theoretical framework

2.1 Concept of Credit Scoring

Nowadays, loan market evolves in big steps. Though everyday new types of banking products are being developed, loan products are still the most demanded. The growing number of loans makes the banks and credit card companies develop tools, which will manage the risk effectively, as risk is the integral part of loans. “One of the most important kits, to classify a bank’s customers, as a part of the credit evaluation process to reduce the current and the expected risk of a customer being bad credit, is credit scoring.” (Abdou, H. & Pointon, J., 2011). Besides, everyday myriad of people apply for different types of loans, which means that it is almost impossible to consider all applications separately.

In the same time customers become more and more demanding. In order to resist the growing competition banks and credit card companies develop new models, which help to satisfy the customer's financial needs. When customers apply for a loan, they want to receive the response of the bank as soon as possible. Credit scoring helps the lenders to automate and accelerate the decision making process. As if a decision to approve loan application takes too long, the customer might look elsewhere for financing.

A credit score is a technique, which helps the lenders to determine if the customer qualify for the current loan product. In other words, credit scoring assesses the creditworthiness of individuals or legal entities. Credit scoring is beneficial for both lenders and customers.

2.2 Historical Background of Credit Scoring

“Credit scoring applications in banking sectors have expanded during the last couple of decades, especially due to the large number of credit applications for different bank products, providing a wide range of new product channels which can be used by these banks.” (Abdou, H. & Pointon, J., 2011). Before that, credit specialists’ judgment was the decisive factor whether the customer received credit or not. Lenders made decisions

based on the behaviour of their previous customers. But that process was rather time-consuming. Also that method was not trustworthy enough, as human mistakes were unavoidable. Gradually, the lenders started to use point systems, that estimated applicant's creditworthiness on the basis of different variables on a customer's credit report. Afterwards, statistical methods were also involved in the decision making process. These methods were based on the information about the current customers. In conjunction with online applications credit scoring models made the decision making process faster and more efficient, than it was before.

Today, many banks both large and small, already have begun to apply the possibilities of credit scoring techniques.

2.3 Credit Scoring Models

Main methods that are used for applying credit scoring are: discriminant analysis, linear regression, probit analysis, logistic regression, Classification and Regression Trees (Decision trees).

The aim of discriminant analysis is to find the discriminant function and to classify objects into one of two or more groups based on a set of features that describe the objects. Decision trees enable to create a tree based classification model, which can be used for prediction. These trees graphically show possible alternatives that enable the lender to choose the most suitable option for the current situation. Other three methods use historical data in order to find the probability of default.

"A newer method of scoring is beginning to be used in the decision-making process. It is based on neural networks consisting of the use of sophisticated technologies and artificial intelligence techniques. The most important feature of neural networks is their ability to learn. Just like human brains, neural networks can learn by samples and dynamically modify themselves to fit the data presented." (Federico Ferretti 2008).

2.4 Upsides of Credit Scoring

The main benefit of credit scoring is that loan officers spend less time for loan approval process. Due to credit scoring, the credit granting process takes only some minutes,

while it could have taken up weeks, or even months in some cases. Another benefit of credit scoring is improved objectivity that helps lenders ensure they are using the same underwriting criteria to all borrowers. This enables lenders to focus on the information that concerns to credit risk and avoid the personal subjectivity of a loan officer. Credit scoring allows the automation of lending process, which in turn gives the opportunity to avoid human mistakes. It greatly reduces the cost of delivering credit. When the loan granting process becomes less costly for lenders, it leads to lower rates overall.

2.5 Limitations of Credit Scoring

Though a lot of upsides of credit scoring, it also has some limitations. One of the main questions is the accuracy of the credit scoring systems for underrepresented groups. Accuracy is very essential consideration for credit scoring. Even if credit scoring decreases the costs of lender, if the models are not enough accurate, additional costs will occur by poorly performed loans.

2.6 Three Main Types of Credit Scoring

There are several types of credit scoring. Each of them has its own characteristics. We are going to give an overview on three main types of credit scoring.

FICO Credit Scores - Although there are a lot of ways of credit scoring calculations, most credit grantors use Classic FICO credit score, which has been in use since 1989. FICO scores are based only on information in customer's credit report and this information is mainly gained from different sources, especially from banks and credit card companies. The three main repositories of these information are Credit Reporting Agencies, which are TransUnion, Experian, and Equifax. It is worth to mention, that Credit Reporting Agencies calculate the score based on the information in their files, while FICO corporation only provides the algorithm for the score calculation.“The FICO scores generally range from 300 to 850 points. Higher FICO scores demonstrate lower credit risk, and lower FICO Scores demonstrate higher credit risk.”(CEB Towergroup, May 2015).

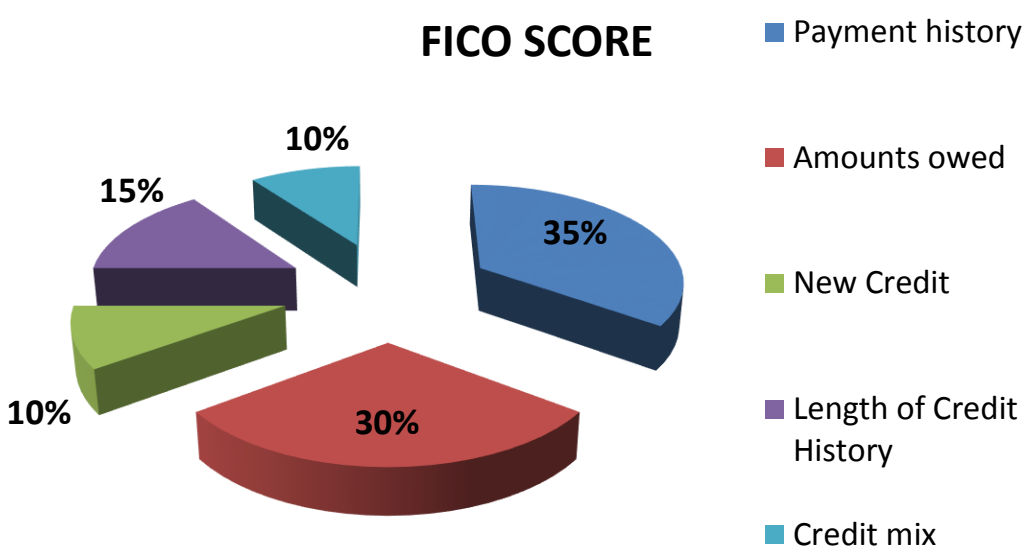
Although FICO has developed other credit scoring models also, the Classic FICO credit score still remains the most widely used score by lenders.

The Vantage score - A new scoring system has been developed by the 3 CRA-s mentioned above, which is called Vantage score. It was created as an alternative to the FICO score, which provides a more steady scoring system for lenders. Besides the loan history, Vantage score also takes into account payment history for rent, utilities, which allows more people to be scored. The Vantage score range runs from 501 to 990.

The PLUS Score - Like all other scores, PLUS score also aims to evaluate the customers capacity of repayment. Unlike other scores, PLUS score is only for customers. It means, that the lenders never make decisions based on PLUS score when granting loans. This characteristics makes this scoring useless. Even if customers have high PLUS score, it is not a guarantee, that their loan application will be approved. The PLUS score range is from 330 to 830.

2.7 FICO Credit Score

"There isn't one standard for credit scores, which is true. But the FICO credit score is probably the longest standing and most widely recognized."(Guina Ryan, 2011). Approximately 90% of lenders use FICO Scores, which help them to make billions of decisions related to credit granting. FICO Scores are calculated from the information in customers credit report. This data is grouped into five categories: payment history, length of credit history, amount owed, new credit, credit mix. How a FICO breaks down is shown in the graph below. (http://www.myfico.com/credit-education/whats-in-your-credit-score/).



These percentages are based on the importance of scoring factors. For some groups of customers, the importance of these scoring factors may be different: for instance, people who have not long credit history, will be factored differently, than those customers, who have longer credit history.

1. **Payment history** - 35% of the total credit score is based on the borrower's payment history. This is one of the decisive factors in FICO score, as every lender wants to know if the borrower paid past debts on time or not. According to FICO score borrower's past long-term behaviour is used to predict future long-term behaviour. FICO indicates, that defaulting on a large loan, will damage the score more strongly than defaulting on a small loan.
2. **Length of credit history** - 15% of the total credit score is based on the borrower's payment history. In general, borrowers with long credit history are considered to be less risky, than borrowers with no credit history. However, even people with short credit history may have high score, depending on the other points of credit report.
3. **Amount owed** - 15 % of the total credit score is based on the amount owed by the borrower. Owing money on credit accounts does not obligatory mean, that the borrower is has low score. However, borrowers,who get close to their credit

limits, are considered overextended. According to FICO it is more likely that this kind of customers will make late or missed payments.

4. **New credit -** New credit determines 10% of a FICO score. FICO considers, that the borrowers, who have loans apply for loans and credit lines at the same time, are in financial trouble. Such behaviour represents greater risk,especially for people, who have short credit history.
5. **Credit mix -** 10 % of the total credit score is based on the credit mix. It is not usually a determining factor. However, borrowers who have different kinds of loans and credit lines and make payments on time are considered less risky. Having good mix of loans and credit lines indicates, that the borrower can handle all sorts of credit.

3. Research Method: Binary logistic regression

3.1 General form of Binary Logistic Regression

Binary logistic regression is regression analysis where the dependent variable is binary. It only contains data coded as 1 or 0. Like other regression models binary logistic regression is also a predictive analysis. The aim of binary logistic regression is to find the model, which describes the relationship between characteristic of interest (dependent variable) and set of independent variables. Before showing how logistic regression general model looks like, let us define odds. Odds of an event are the ratio of the probability that an event will occur to the probability that it will not occur. If the probability of presence of the characteristic of interest is p, the probability of absence of the characteristic of interest is 1-p. Then the corresponding odds is a value given by this formula:

$$odds = \frac{p}{1-p} = \frac{probability\ of\ presence\ of\ characteristic}{probability\ of\ absence\ of\ characteristic}$$

Since logistic regression calculates the probability of an event occurring over the probability of an event not occurring, the influence of independent variables is usually explained in terms of odds. With logistic regression the mean of dependent variable p in terms of dependent variable x is given by the equation $p=\alpha + \beta x$. This is not a good model, as values of $\alpha + \beta x$ does not fall between 0 and 1. Logistic regression gives a solution to this problem by transforming the odds using the natural logarithm. With logistic regression we model the natural log odds as a linear function of the independent variable:

$$logit(y) = ln(odds) = ln(\frac{p}{1-p}) = \alpha + \beta x \qquad (1)$$

where p is the probability of interested outcome and x is the independent variable. The parameters of logistic regression are α and β. This is the simple logistic model.

From equation (1) we can derive an equation for the prediction of the probability as

$$p = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} = \frac{1}{1+e^{-(\alpha+\beta x)}} \qquad (2)$$

Assuming that we have multiple predictors we may construct a general logistic model as

$$logit(y) = ln(\frac{p}{1-p}) = \alpha+\beta_1 x_1 + \cdots + \beta_k x_k \qquad (3)$$

From equation (3) we can derive an equation for the prediction of the probability as

$$p = \frac{e^{\alpha+\beta_1 x_1+\cdots+\beta_k x_k}}{1+e^{\alpha+\beta_1 x_1+\cdots+\beta_k x_k}} = \frac{1}{1+e^{-(\alpha+\beta_1 x_1+\cdots+\beta_k x_k)}} \qquad (4)$$

3.2 Logistic Curve

When the dependent variable is binary and independent variable is numerical, logistic model fits a logistic curve to the relationship between x and y. Logistic curve is a common "S" shape (sigmoid curve).

A simple logistic function is defined by the following formula

$$y = \frac{e^x}{1+e^x}.$$

This equation can be extended to the form

$$y = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} = \frac{1}{1+e^{-(\alpha+\beta x)}}$$

which is graphed in Figure 1.

$\frac{e^x}{1+e^x}$

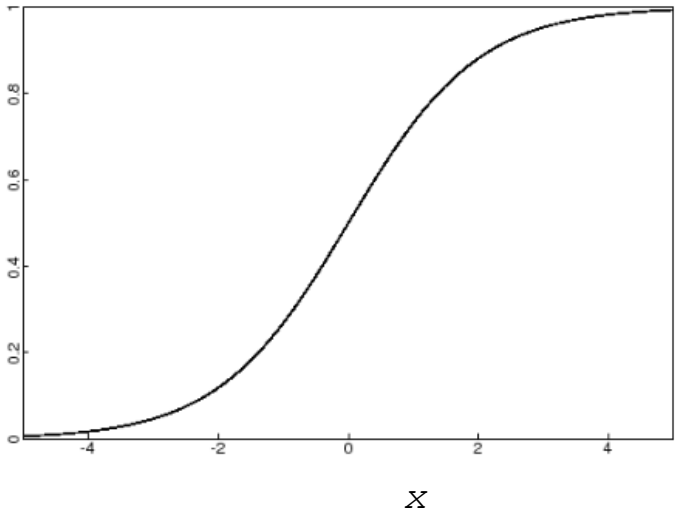


Figure 1.Logistic Curve

Figure 1 shows logistic function, where α is 0 and β is 1.

3.3 Assumptions of Binary Logistic Regression

Unlike general linear models, binary logistic regression does not have many key assumptions, particularly it does not require a linear relationship between the dependent and independent variables, normality of the error distribution, homoscedasticity of the errors and measurement level of the independent variables.

(http://www.statisticssolutions.com/assumptions-of-logistic-regression/)

However logistic regression still requires other assumptions.

1. Binary logistic regression requires the dependent variables to be binary.

2. Since binary logistic regression assumes that that P(Y=1) is the probability of event occurring, it requires that the dependent variable is coded accordingly.

3. Model should be fitted correctly. It means that all meaningful variables should be included. Also, it should not be over fitted with meaningless variables included.

4. Binary logistic regression requires each observation to be independent. Also, it should have little or no multicollinearity, which means that independent variables are not linear functions of each other.

5. Binary logistic regression requires linearity of the relationship between independent variables and log odds. Meanwhile, it does not require a linear relationship between dependent and independent variables.
6. Binary logistic regression requires quite large sample sizes. Studies with small sample sizes overestimate the effect measure. Also the more independent variables are included in the model, the larger sample size is required.

3.4 Maximum Likelihood Estimation

Although logistic regression model looks like simple linear regression model, the underlying distribution is binomial, and α and β parameters cannot be estimated in the same way as for simple linear regression. The coefficients are usually estimated by the Maximum Likelihood Model (Park, Hyeoun-Ae, April 2013). The likelihood is a probability to get observed values of the dependent variable given the observed values of independent variables. The likelihood varies from 0 to 1 like any other probabilities.
We can write

$$P(Y=y_i) = {P_i}^{1-y_i}(1-P_i)^{y_i}$$

where P_i is the probability of the i-th observation, y_i is the value of random variable Y that takes value 0 or 1. Assuming that our n observations are independent the likelihood of the data is equal to

$$L=\prod_{i=1}^{n} {P_i}^{1-y_i}(1-P_i)^{y_i}$$

Maximum Likelihood method will provide values for α and β which maximise L function.

3.5 Evaluation of Binary Logistic Regression Model

There are several ways for estimating logistic regression. Firstly, the overall model should be evaluated. Secondly, the significance of every explanatory variable needs to be assessed. Thirdly, the predictive accuracy needs to be evaluated.

Overall model evaluation

a) Likelihood ratio test

Due to overall model evaluation we can see how strong the relationship between all independent variables and dependent variable is. If logistic regression with k independent variables demonstrates an improvement over the model without independent variables (null model), then it provides a better fit to data (Park, Hyeoun-Ae, April 2013). This is performed using the likelihood ratio test, which compares the likelihood of the data under the full model with the likelihood of the data under the model without independent variables. The overall fit of the model with k coefficients can be assessed via likelihood ratio test which tests the null hypothesis

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_k$$

-2 log likelihood of the null method is compared with 2 log likelihood of the given model. Likelihood of null method is the likelihood of obtaining the observation if explanatory variables have no impact on the outcome. Likelihood of the given model is likelihood of obtaining the observation if all explanatory variables are included in the model.

The difference of these 2 indicates a goodness of fit index G, χ^2statistic with k degrees of freedom. It measures how well independent variables influence on the dependent variable.

$$G = \chi^2 = \textit{(-2 log likelihood of null model) - (-2 log likelihood of the given model)}$$

If the p-value for the overall model fit statistic is less than 0,005, then decline H_0 with the conclusion that at least one of the independent variables has impact on the outcome or dependent variable.

b) Chi-square Goodness of Fit Tests

Chi-square goodness of fit test is a non parametric test that is used to find out how the observed value of a given event is significantly different from the expected value. The hypothesis for Chi-square goodness of fit test is as follows.

Null hypothesis: In Chi-square goodness of fit test, the null hypothesis assumes that there is no significant difference between the observed and expected value.

Alternative hypothesis: In Chi-square goodness of fit test, the alternative hypothesis assumes that there is significant difference between the observed and expected value. If the p-value is less than significance level, the null hypothesis is rejected.

In linear regression residuals are defined as y_i-$\hat{y}_i$ where y_i is the observed value of the variable for i-th subject, and $\hat{y}_i$ is the predicted value for i-th subject. For logistic regression, where y_i is 1 or 0, the corresponding prediction from the model is as

$$\hat{y}_i = \frac{exp(\alpha+\beta_1 x_{i1}+\cdots+\beta_k x_{ik})}{1+exp(\alpha+\beta_1 x_{i1}+\cdots+\beta_k x_{ik})}$$

Chi-square test is based on residuals y_i-$\hat{y}_i$. A standardized residual is defined as

$$r_i = \frac{y_i-\hat{y}_i}{\sqrt{\hat{y}_i(1-\widehat{y-}_i)}},$$

and χ^2 statistic can be formed as

$$\chi^2 = \sum_{i=1}^{n} {r_i}^2$$

This statistic follows χ^2 distribution with n-(k+1) degrees of freedom.

c) Hosmer-Lemeshow test

Hosmer-Lemeshow test also measures how good the model is. The test evaluates whether observed event rates match expected event rates in subgroups of the model

population. Hosmer-Lemeshow test is implemented by dividing the predicted probabilities into ten equal groups, according to their values (deciles).

The hypotheses is as follows

H_0: Actual and predicted event rates are similar across 10 deciles.

H_1: They are not the same.

The value of the test statistics is

$$\chi^2 = \sum_{g=1}^{10} \frac{(O_g - E_g)^2}{E_g}$$

where O_g are the observed events, and E_g are the expected events for the g-th risk decile group. The test statistic asymptotically follows a χ^2 distribution with 8 (number of groups-2) degrees of freedom. Small values with large p-value closer to 1 means a good fit to the data. Large values with $p < 0,05$ means a poor fit to the data.

Statistical significance of individual regression coefficients

After evaluating the overall model, the next step is to assess the significance of every independent variable. The coefficient of i-th explanatory variable indicates the change in the predicted log odds for one unit change in the i-th explanatory variable, when all other explanatory variables remain unchanged.

a) Likelihood ratio test

As we mentioned above, the likelihood ratio test is used to evaluate the overall fit model. The test is also used to evaluate statistical significance of individual predictors. The likelihood ratio test for particular parameter compares the likelihood of obtaining the data when the parameter is 0 (L_0) with the likelihood (L_1) of obtaining the data evaluated at the MLE of the parameter.

The test statistic is calculated as

$$G=-2ln\frac{L_0}{L_1}=-2(lnL_0 - lnL_1)$$

This statistics is compared with χ^2 distribution with 1 degree of freedom.

b) Wald statistic

The Wald statistic is used to test the significance of individual coefficients in a given model (Bewick et al.,2005). The statistic is the ratio of the square of the regression coefficient to the square of standard error of the coefficient. The calculation is as follows

$$W_j = (\frac{coefficient}{SEcoefficient})^2$$

Each Wald statistic is compared with a χ^2 distribution with 1 degree of freedom. The calculation of Wald statistic is easy. However, the reliability of the test is questionable, particularly for small samples. For data that produces large estimates of coefficient, the standard error is often inflated, which in turn results in a lower Wald statistic. Consequently, explanatory variable may be incorrectly assumed as insignificant in the model.

Predictive Accuracy and Discrimination

a) Classification table

The classification table is a tool to assess the predictive accuracy of logistic regression model. In the below table the observed values for the dependent outcome and the predicted values are cross-classified. For instance, if the cut off value is 0,5, all predicted values above 0,5 can be classified as predicting an event, and all predicted values below 0,5 as not predicting the event.

Table 1. Sample Classification Table

Observed	Predicted	
	0	1
0	a	b
1	c	d

where a and d are the number of cases that are predicted correctly, b and c are the number of cases that are not predicted correctly.

If we see many counts in the a and d cells, and few in the b and c cells, we can conclude that our model has a good fit. Two indicators are used to evaluate the accuracy of a test

that predicts binary outcome-sensitivity and specificity. Sensitivity is the proportion of true positives (Y=1). Specificity is the proportion of true negatives (Y=0). Sensitivity is calculated by the formula d/c+d. Specificity is calculated by the formula a/a+b. The values of sensitivity and specificity depend on the cut off value we choose. If we increase the cut point, fewer observations will be predicted as positive. In other words, fewer of Y=1 observations will be predicted as positive, which means that the sensitivity will decrease. While, more of the Y=0 observations will be predicted as negative, which means that the specificity will increase. The model is perfect if it has 100% sensitivity and 100% specificity. In practice this result is not usually attainable.

b) ROC curve (Receiver Operating Characteristics)

The ROC curve is a fundamental technique for diagnostic test evaluation. It extends the above two-by-two table and examines the full range of cutoff values from 0 to 1. For each possible cutoff value, a two-by-two table can be formed. The ROC curve is a plot of the values of sensitivity versus one minus specificity, as the value of cut off is increased from 0 to 1. The ROC curve is more informative than the classification table because it reflects the predictive power for all cut off values.

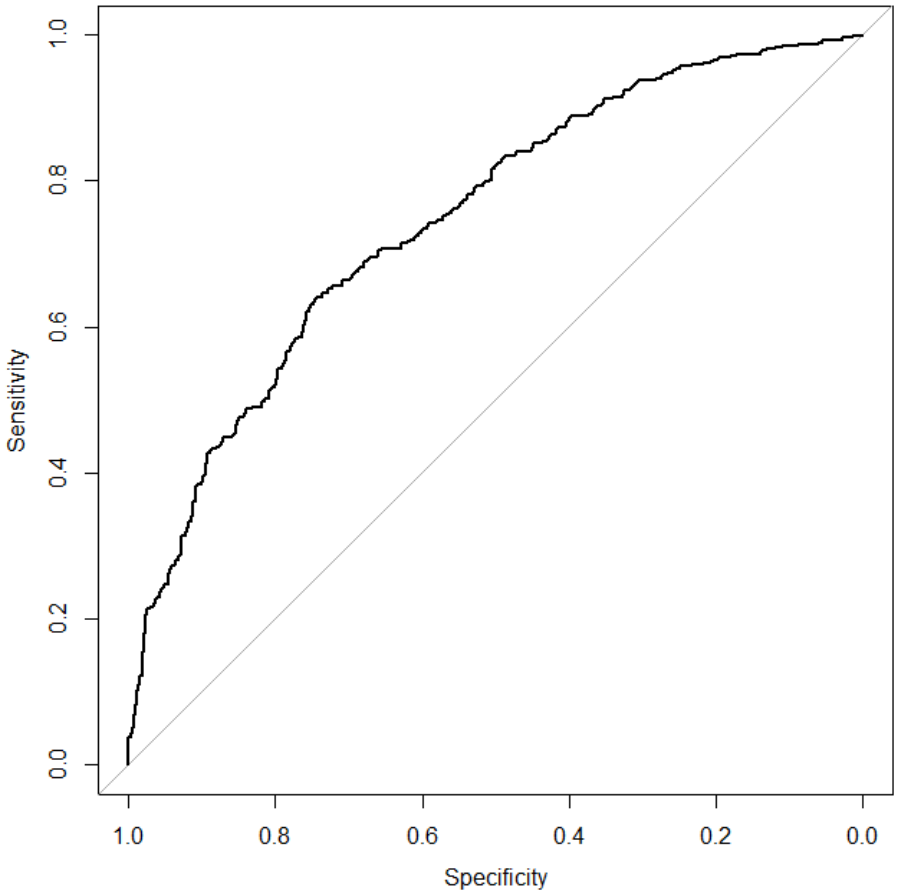


Figure 2. An example of ROC Curve

Area under the ROC curve is a way of summarizing the discrimination ability of a model. An area of 1 indicates a perfect test. While an area of 0.5 indicates a worthless test. An area over 0.7 is considered very good model.

4. Application of Logistic Regression

4.1 Data Collection and Preparation

At our disposal there is information about 3985 loan applicants. The database includes two groups of applicants: "good" customers (Y=1) who repaid their loan (event occurred), and "bad'customers (Y=0), who defaulted on their loans. 29% of applicants are bad customers, while 71% of applicants are good customers.

Table 2. Summary of Database

Number of applicants	3985
Number of attributes	10

Each applicant is described by 11 variables, shown in Table 3. Among 11 variables there are 10 dependent variables and 1 dependent variable. The response variable Y is binary, which takes values 0 or 1.
7 independent variables are "Scale" variables and three variables are "Ordinal" variables.

Table 3. Variables used in the model

Variable	**Description**	**Type**
Y	Applicant status	Ordinal
V1	Gender	Ordinal
V2	Age	Scale
V3	Loan period in days	Scale
V4	Monthly income in EUR	Scale
V5	Monthly outcome in EUR	Scale

V6	Marital status	Ordinal
V7	Education	Ordinal
V8	Number of children	Scale
V9	Number of real estate units	Scale
V10	Number of payment problems	Scale

The values that ordinal variables take is given in Table 4.

Table 4

	0	**1**
Applicant status	"Bad" customers	"Good" customers
Gender	Male	Female
Marital Status	Unmarried	Married
Education	Uneducated	Educated

Frequency of ordinal variables is given in Table 5.

Table 5

		Frequency	**Percent**
Applicant status	"Bad" customers	1160	29.1%
	"Good" customers	2825	70.9%
Gender	Male	1946	48.8%
	Female	2039	51.2%
Marital Status	Not married	2817	70.7%
	Married	1168	29.3%
Education	Uneducated	3258	81.8%
	Educated	727	18.2%

Descriptive statistics of scale variables is given in Table 6.

Table 6. Descriptive statistics

	Minimum	Maximum	Mean	St.Deviation
Age	19	68	40.6	12.1
Loan period in days	1	720	111.6	121.2
Monthly income	95	14004	793.7	507.2
Monthly outcome	4	5000	328.9	250.4
Number of children	0	10	0.6	0.9
Number of real estate units	0	8	0.6	0.8
Number of payment problems	0	42	1.3	2.5

After the data collection, data preparation is also essential part of study. When working with a real database we should take into account that some data might be missing.

4.2 Binary Logistic Regression with All Independent Variables

In this thesis the IBM SPSS software was used to conduct logistic regression. Let us see what happened when we used all 10 explanatory variables as predictors in our model. After processing binary logistic regression, statistical outputs will be generated. Based on the "Case Processing Summary" output it is visible that 3983 cases used out of 3985. It is explained by the fact, that two cases included missing data.

Table 7. Case Processing Summary

Unweighted Cases		N	Percent
Selected Cases	Included in Analysis	3983	99,9
	Missing Cases	2	,1
	Total	3985	100,0
Unselected Cases		0	,0
Total		3985	100,0

Firstly, we will try to implement the overall model evaluation.

Table 8. Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	4464,785	,082	,117

In this summary it is visible that -2 Log likelihood is 4464.785. By itself this number is not very informative. The p-value for our overall model is 0.000 (less than 0.05), which means that null hypothesis is rejected and there is evidence that at least one of the explanatory variables contributes to the prediction of the outcome.

Cox & Snell R square and Nagelkerke R square are both methods of calculating the explained variation. For our model the explained variation ranges from 0.082 to 0.117 depending on whether we reference Cox & Snell R square or Nagelkerke R square, respectively. Nagelkerke R square is the modification of Cox & Snell R square and is more preferable to use.

The next test that assesses the goodness of fit of a statistical model is Hosmer-Lameshow test.

Table 9. Contingency Table for Hosmer and Lemeshow test

		Y = 0		Y = 1		
		Observed	Expected	Observed	Expected	Total
Step 1	1	199	199,709	199	198,291	398
	2	172	175,277	226	222,723	398
	3	156	158,986	242	239,014	398
	4	145	141,708	253	256,292	398
	5	129	122,820	269	275,180	398
	6	120	105,548	278	292,452	398
	7	95	90,004	303	307,996	398
	8	63	73,959	335	324,041	398
	9	50	57,731	348	340,269	398
	10	31	34,259	370	366,741	401

Table 9 shows that observed proportions of events are rather similar to the predicted probabilities of occurrence in 10 subgroups.

In order to decide whether the differences can be explained by chance only, we perform Hosmer-Lemeshow chi-square test. Based on Table 10 table we can see that p-value is 0.497, which is more than 0.05. This value indicates that we fail to reject null hypothesis, which means that actual and predicted event rates are similar across 10 deciles.

Table 10. Hosmer and Lemeshow test

Step	Chi-square	df	Sig.
1	7,372	8	,497

After overall model evaluation we analyze how important each of the variables is. The "Variables in the Equation" table shows the contribution of each independent variable to the model. Also, the output shows, if the explanatory variables are significant or not. This table is shown below.

Table 11. Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	V1(1)	-,335	,080	17,463	1	,000	,715
	V2	,020	,003	36,819	1	,000	1,021
	V3	-,001	,000	2,812	1	,094	,999
	V4	,000	,000	,165	1	,685	1,000
	V5	,001	,000	8,999	1	,003	1,001
	V6(1)	-,134	,088	2,310	1	,129	,875
	V7(1)	-,411	,109	14,145	1	,000	,663
	V8	-,139	,043	10,488	1	,001	,870
	V9	,568	,061	86,047	1	,000	1,764
	V10	-,063	,014	19,543	1	,000	,939
	Constant	,472	,215	4,829	1	,028	1,603

V1, V2,....,V10 are the independent variables included in the binary logistic model. Constant is the expected value of log-odds of dependent variable when all of the predictor variables equal zero.

B (beta coefficients) are the values for the logistic regression equation for predicting the response variable from explanatory variables.
For our model the prediction equation is as follows

$$log(p/1-p)=0.472-0.335*V1+0.020*V2-0.01*V3+0.000*V4+0.001*V5-0.134*V6-0.411*V7-0.139*V8+0.568*V9-0.063*V10$$

Beta coefficients show the amount of change expected in the log odds when there is a one unit change in the predictor variable holding all other predictors constant. For the independent variables that are not significant the coefficients do not significantly differ from 0. Because these coefficients are in log odds units, they are often difficult to interpret, and converted into odd ratios. These values are shown in "Exp (B)" column.
"S.E"-s are standard errors associated with the coefficients. The standard error is used to test whether the parameter is significantly different from 0 or not. Standard errors are also used in the calculation of Wald statistic. Also, they can be used to form a confidence level for the parameter.
"Wald" tests the hypothesis that the constant equals 0. For our model this hypothesis is rejected because the p-value, which is listed in the "Sig" column is less than the critical p-value (0.05 or 0.01). Therefore, we can conclude that the constant is not 0.
"Df" is the degree of freedom for the Wald chi-square. There is given 1 degree of freedom for each predictor in the model.
"Exp(B)"-s are the exponentiations of the beta coefficients, which are the odds ratios of the predictors. The odds ratio represents that an outcome will occur given a particular property, compared to the odds of the outcome occurring in the absence of that property. As we mentioned above, the prediction equation is given in log odds. Taking e to the power for both sides of the equation we can express the expression in odds.
"Sig." is p-value of significance test of beta. Usually the coefficients which p-values are less than 0,05, are considered to be significant. Based on our output we can see that all of the explanatory variables are significant except V3 (Loan period), V4 (Monthly income), V6 (Marital status). The p-values of these coefficients are greater than 0.05.
After evaluating the statistical significance of individual coefficients, let us evaluate the predictive accuracy and discrimination of the model. Based on the "Classification table" output we assess the predictive accuracy of the model. IBM SPSS sets cutoff value 0.5 as default. In our analysis we used another cutoff values as well: 0.4, 0.6, 0.7. We also

calculated the sensitivity and specificity for these values of cut off. The classification table, where the cut-off value is 0.5 is shown below.

Table 12. Classification Table (Cutoff value is 0.5)

Observed			Predicted		
			Y		Percentage Correct
			0	1	
Step 1	Y	0	66	1094	5,7
		1	77	2746	97,3
		Overall Percentage			70,6

66 is the number of cases that were both predicted and observed as 0. 1094 is the number of cases that were observed as 0, but were predicted as 1. 77 is the number of cases that were observed as 1, but were predicted as 0. 2746 is the number of cases that were both predicted and observed as 1. In classification table overall percentage gives the information that 70.6% of cases were correctly predicted.
Next, the classification tables for different cutoff values are introduced.

Table 13. Classification Table (Cutoff value is 0.4)

Observed			Predicted		
			Y		Percentage Correct
			0	1	
Step 1	Y	0	7	1153	,6
		1	7	2816	99,8
		Overall Percentage			70,9

Table 14. Classification Table (Cutoff value is 0.6)

Observed			Predicted		
			Y		Percentage Correct
			0	1	
Step 1	Y	0	449	711	38,7
		1	540	2283	80,9
		Overall Percentage			68,6

Table 15. Classification Table (Cutoff value is 0.7)

Observed			Predicted		
			Y		Percentage Correct
			0	1	
Step 1	Y	0	766	394	66,0
		1	1104	1719	60,9
		Overall Percentage			62,4

In the next table sensitivity and specificity have been calculated for different values of cutoff.

Table 16

Cutoff value	Sensitivity	Specificity
0.4	99 %	6%
0.5	97 %	6 %
0.6	80 %	39%
0.7	60%	66%

Table 13 shows that when we increase the cutoff values, the sensitivity decreases, while the specificity increases.

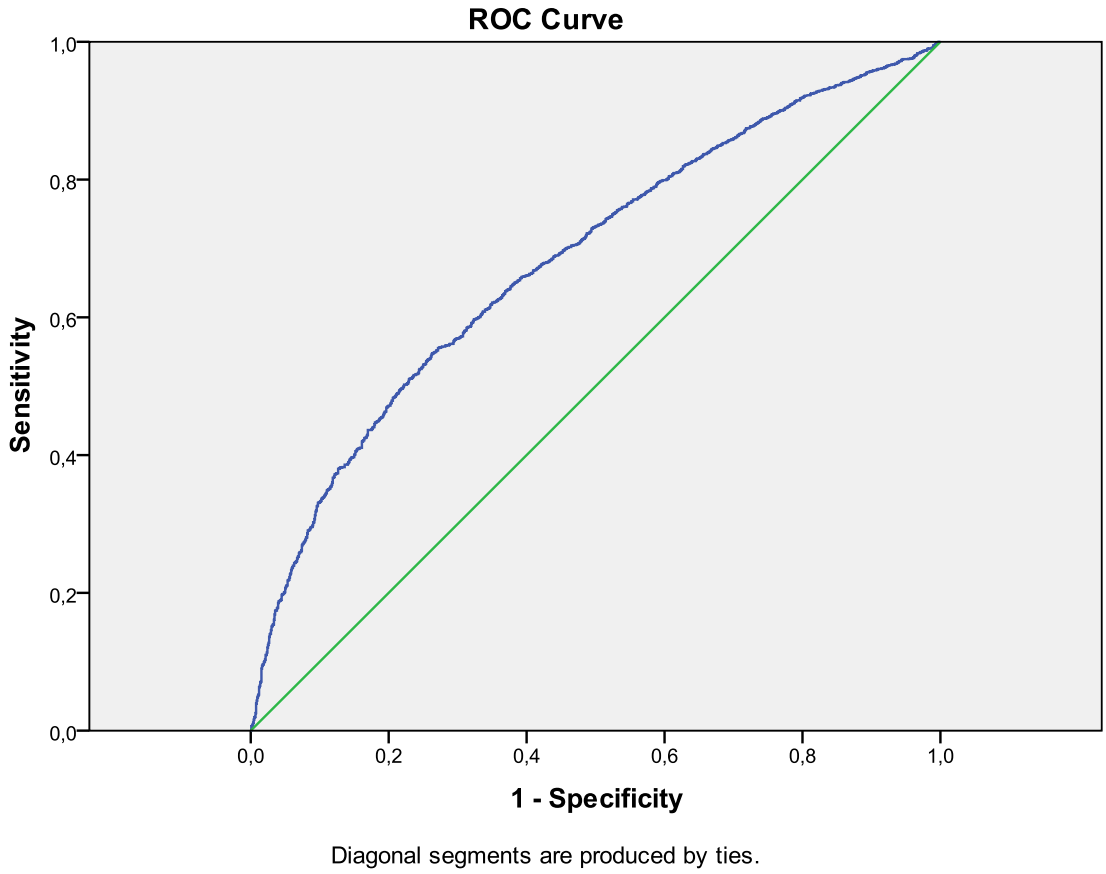


Figure 3. ROC curve

Area under the Roc curve (see Table 17) is 0.683 with 95% confidence level. Also, the area under the curve is significantly different from 0.5 since p-value is 0.000 meaning that logistic regression classifies the group significantly better than by chance.

Table 17. Area Under the ROC curve

			Asymptotic 95% Confidence Interval	
Area	Std. Error	Asymptotic Sig.	Lower Bound	Upper Bound
,683	,009	,000	,666	,701

4.3. Logistic Regression With Selected Independent Variables.

Here we eliminate statistically insignificant variables from the model. Based on the Table 8, we got that the V3 (Loan period), V4 (Monthly income), V6 (Marital status) variables were not significant. Next we implement the same steps as in the last subchapter, but eliminating these variables from the model.

Like in the previous subchapter, let us firstly evaluate the overall model. "Model Summary" is as follows.

Table 18. Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
	4470,063	,081	,115

It is visible that -2 log likelihood is 4470.063, while it was 4464.785 in the full model with all 10 variables. The value of Nagelkerke R square is 0.115, which means weaker predictive capacity than before, as for the full model Nagelkerke R square was 0.117.

Table 19. Contingency Table for Hosmer and Lemeshow test

		Y = 0		Y = 1		
		Observed	Expected	Observed	Expected	Total
Step 1	1	192	198,845	206	199,155	398
	2	178	174,523	220	223,477	398
	3	156	158,890	242	239,110	398
	4	144	141,536	254	256,464	398
	5	133	122,377	265	275,623	398
	6	115	106,268	283	291,732	398
	7	93	90,661	305	307,339	398
	8	68	74,377	331	324,623	399
	9	51	58,197	347	339,803	398
	10	30	34,324	370	365,676	400

Table 19 shows that observed proportions of events are rather similar to predicted probabilities of occurrence in 10 deciles.

Table 20. Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	5,447	8	,709

"Hosmer and Lemeshow Test" shows that in this case also we fail to reject the null hypothesis, as p-value is 0.709 and again is larger than 0.05.

Next, we evaluate the significance of independent variables. In our model we included V1 (Gender), V2 (Age), V5(Monthly outcome), V7 (Education), V8 (Number of children), V9 (Number of real estate units), V10 (Number of payment problems) variables.

Using 7 independent variables in our model, we got the following result (see table 19).

Table 21. Variables in the equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	V1(1)	-,325	,079	16,984	1	,000	,723
	V2	,021	,003	42,267	1	,000	1,022
	V5	,001	,000	10,426	1	,001	1,001
	V7(1)	-,415	,109	14,586	1	,000	,660
	V8	-,128	,041	9,472	1	,002	,880
	V9	,569	,061	87,496	1	,000	1,767
	V10	-,063	,014	19,749	1	,000	,939
	Constant	,262	,184	2,035	1	,154	1,300

Table 21 shows that all of the explanatory variables are significant, as p-values for all of them are larger than 0.05.

The classification table below shows the predictive accuracy of the selected variable model, when cutoff value is 0.5.

Table 22.Classification table (Cutoff value is 0.5)

Observed			Predicted		
			Y		Percentage Correct
			0	1	
Step 1	Y	0	60	1100	5,2
		1	77	2746	97,3
		Overall Percentage			70,4

Table 22 gives us information that 5.2% of "bad" applicants were correctly classified and 97.3% of "good" applicants were correctly classified. The predictive accuracy for overall model is 70.4%.

Next the classification tables for different cutoff values are given.

Table 23. Classification table (cut off value is 0.4)

Observed			Predicted		
			Y		Percentage Correct
			0	1	
Step 1	Y	0	6	1154	,5
		1	6	2817	99,8
		Overall Percentage			70,9

Table 24. Classification table (Cutoff value is 0.6)

Observed			Predicted		
			Y		Percentage Correct
			0	1	
Step 1	Y	0	451	709	38,9
		1	537	2286	81,0
		Overall Percentage			68,7

Table 25. Classification table (cut off value is 0.7)

Observed			Predicted		
			Y		Percentage
			0	1	Correct
Step 1	Y	0	766	394	66,0
		1	1093	1730	61,3
		Overall Percentage			62,7

Like in the full model, in selected variables model also, the predictive accuracy for 0.4 value of cutoff is higher than for other values of cutoff.
In this case also we calculated sensitivity and specificity for different cutoff values.

Table 26

Cutoff value	Sensitivity	Specificity
0.4	99%	0.5%
0.5	97%	5%
0.6	81%	39%
0.7	61%	66%

Based on the "Sensirivity" and "Specificity" values for different cut offs we can conclude that 0.7 value of cut off is more preferable than others. For other cut off values that we analyzed we see that high values of sensitivity are combined with low values of specificity, from which we can conclude that 0.7 value of cut ff gives more balanced results.

For selected variables model ROC curve looks like follows.

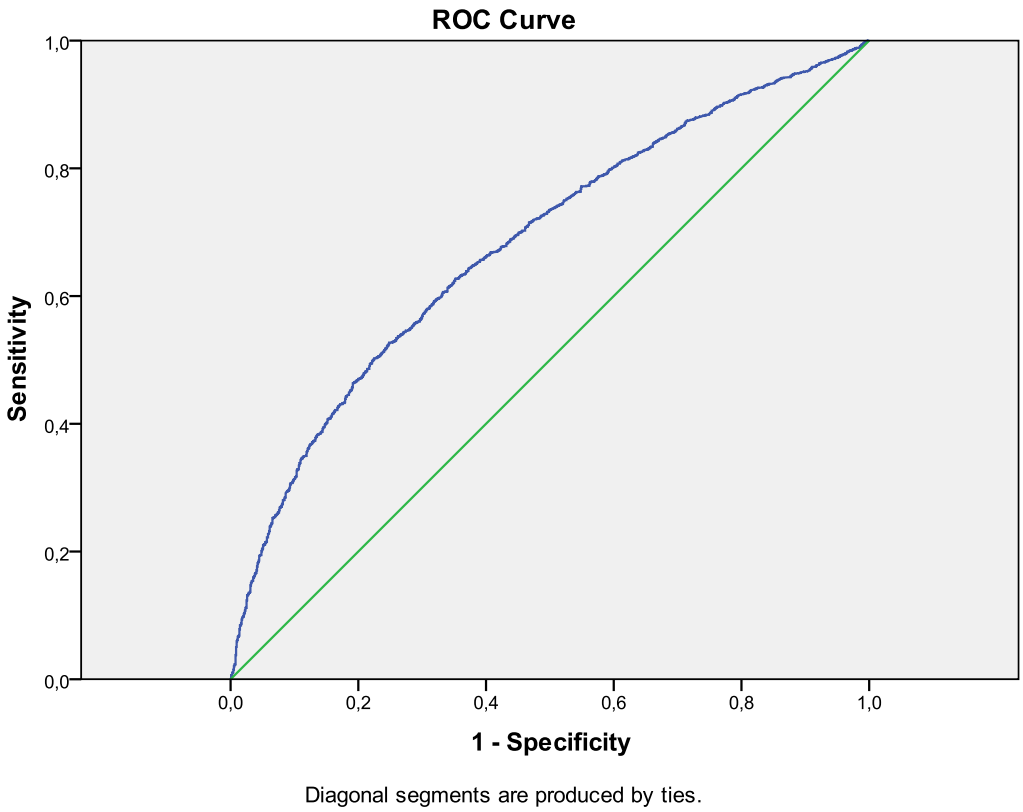


Figure 4. ROC curve

The area under the curve is 0.682 now with 95% confidence interval which is slightly less than in case of full model. (See Table 27)

Table 27. Area under the curve

			Asymptotic 95% Confidence Interval	
Area	Std. Error	Asymptotic Sig	Lower Bound	Upper Bound
,682	,009	,000	,665	,700

We can see that in general new model with 7 independent variables has almost the same quality as the full model with 10 variables.

Conclusion

In this Master's thesis we used real data collected from 3985 loan applicants both "good" and "bad". Our dependent variable took 2 values: "0" or "1" depending on whether the applicant was bad or good customer. 10 explanatory variables were included in our model. 7 of them were scale variables and 3 were ordinal variables. We conducted binary logistic regression in IBM SPSS software, which calculated the predicted probability of the event. We excluded all three non significant variables from the model. By using the final model, 70.4% of the cases were correctly classified in the case of cutoff value 0.5. We also calculated ROC curve of our model and got the value of "Area under the curve" indicator, which was 0.682. Such a value of AUC is usually considered as reasonably good.

To sum up, logistic regression is a powerful tool which helps to accelerate decision making process. With logistic regression lenders can implement the loan approval process more accurately.

References

1. Ahmet Burak Emel, Muhittin Oral, Arnold Reisman, Reha Yolalan (2003). A credit scoring approach for the commercial banking sector.
2. Abdou, H. & Pointon, J. (2011) Credit scoring, statistical techniques and evaluation criteria
3. Mark Schreiner (2004). Benefits and Pitfalls of Statistical Credit Scoring for Microfinance
4. http://www.investopedia.com/terms/f/ficoscore.asp
5. Guina Ryan, (2011). Examining different types of credit scoring
6. Abdou, H. & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria
7. Gouvêa Maria Aparecida, Gonçalves Eric Bacconi, (2007) Credit Risk Analysis Applying Logistic Regression, Neural Networks and GeneticAlgorithms Models.
8. Dean Caire and Robert Kossmann, (2003). Credit Scoring: Is It Right for Your Bank?
9. Park, Hyeoun-Ae (2013), An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain
10. Federico Ferretti (2008), The Law and Consumer Credit Information in the Europenian Community.
11. CEB Towergroup (May 2015),Understanding FICO scores.
12. http://www.statisticssolutions.com/assumptions-of-logistic-regression/
13. http://www.myfico.com/credit-education/whats-in-your-credit-score/

Non-exclusive licence to reproduce thesis and make thesis public

I, Nare Torosyan,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

1.1. reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

1.2. make available to the public via the university's web environment, including via the DSpace digital archives, as of **16.05.2017** until expiry of the term of validity of the copyright, Application of Binary Logistic Regression in Credit Scoring,

supervised by Prof. Kalev Pärna,

2. I am aware of the fact that the author retains these rights.

3. This is to certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu/Tallinn/Narva/Pärnu/Viljandi, **16.05.2017**