UNIVERSITY OF TARTU

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

Institute of Mathematical Statistics

Salome Tabagari

Credit scoring by logistic regression

Master's Thesis

In Financial and Actuarial Mathematics

Supervisor: Prof. Kalev Pärna

Tartu 2015

I have written the Master Thesis myself, independently. All of the other authors' texts, main viewpoints and all data from other resources have been referred to.

Author : .....

(*signature*)

13.05.2015

(*date*)

Table of Contents

Credit Scoring By Logistic Regression

Abstract

Today banking business' most successful products are loans and credits given to the clients. In order to make a decision whether to accept or reject a loan application banks gather information from applicants. In the past, decision was made by individual bank's expert. It was not efficient way for banks, because competition was growing, thus they introduced better method – credit scoring. Credit scoring is one of the most effective and successful methods in finance and banking. With help of credit scoring methodology it is easier to make correct and fast decisions.

An overview of the credit scoring is given in the following thesis. A real data set is used to demonstrate how to calculate applicants' scores. For this purpose one of the most frequently used statistical method- logistic regression – is used.

Keywords: credit scoring, logistic regression.

Laenutaotluste hindamine logistilise regressiooni abil

Lühikokkuvõte

Pankade kõige tulusam toode on laenud. Selleks, et otsustada, kas laenu anda või mitte, kogub pank laenutaotlejalt informatsiooni. Vanasti tegi laenu andmise otsuse panga ekspert. Kuna konkurents laenuturul kasvas, siis see meetod muutus ebaefektiivseks ja pangad võtsid kasutusele laenutaotluste hindamise kvantitatiivse metoodika (ingl.k. *credit scoring*).

Käesolevas magistritöös antakse esmalt ülevaade laenutaotluste hindamise erinevatest meetoditest. Lähemalt vaadeldakse üht kõige levinumat sellekohast statistilist meetodit – logistilist regressiooni. Seejärel rakendatakse logistilist regressiooni reaalsel laenutaotlejate andmestikul.

Märksõnad: laenutaotluste hindamine, logistiline regressioon

1.Introduction

Problem Statement: Credit score has very important role in financial institutions' working. Using credit scoring method, credit officers easily can predict probability of default to avoid high losses. Some banks are still using judgmental decision means credit officers have interview with applicant and analyze gathered information separately. But most of banks try to use easier and useful method, such as credit scoring. Using method depends on type of the credit or loan. For small credits, like credit card or consumer credit, lenders prefer to use credit scoring. The system makes decision based on the information that is known from the previous customer's database.

Professional lenders try to use more effective methods for credit scoring to make more accurate decisions. The methods are not always correct, some good applicant can get bad score. Relatively long practice of credit scoring shows that one of the most useful and efficient methods in credit scoring is classical logistic regression that depends on customers' historical data.

Purpose: The primary purpose of this Master's thesis is show importance of credit scoring for lenders, to find variables that are more frequently used in credit scoring system. The secondary purpose is to how logistic regression works and compare the logistic regression results with actual results, in order to see how correct the decision rule is.

Research Method: This is a quantitative research mostly. We collect historical data from customers to define their characteristics. Research method is to be observed by the statistical methods and logistic regression. We follow the general research principles: "All aspects of the research are carefully designed before data collection procedure. And the analysis is targeted for the precise measurements" (Islam, Zhou, Li 2009).

Findings: First we have shown that logistic regression can be successfully applied to solve credit scoring problem. Secondly, we have identified predictor variables - the features of loan applicants- which play important role in dividing applicants between two classes – "good" and "bad".

The Structure of the thesis is as follows. In Chapter 2 we give a historical overview of credit scoring and describe some well-known credit scoring systems (FICO, in particular). This description is done in rather general terms since for the confidentiality reasons the details of main credit scoring systems are not available. In Chapter 3 we describe our study design, data collection procedures, and basic data characteristics. In Chapter 4 a brief introduction into logistic regression is given, followed by its application on real customer's data. All statistical calculations are performed by means of IBM SPSS software.

2.Theoretical Basis

2.1.What Is Credit Scoring?

Credit is very important product in banking and financial institutions. There is always a customer who needs loan to buy different stuffs. Loans are always accompanied by risks. The risk for financial institutions depends on how well they can separate good applicants from bad applicants. For solving this problem, lenders started using "credit scoring".

Credit scoring is a method for defining the risk of loan applicants. By calculating credit score lenders can make decision who gets credit, how good creditor he/she could be, what will the percent and how much credit or loan they can get.

"A lender commonly makes two types of decisions: first, whether to grant credit to a new application or not, and second, how to deal with existing application, including whether to increase their credit limits or not." (Thomas, Edelman and Crook 2002, p1)

Lenders use "historical" data gathered from observed of applicant to build applicants scorecard. They gather data about applicants, such as applicant's income, financial asset, type of work, working current place, residual status, time with bank, credit history, if he/she had default or problem with payment.

Scoring model is better way for decision making than traditional judgmental method, but the model is not perfect-sometimes a bad applicants will receive high score and will be accepted, and vice versa, a good applicant can get low score and be rejected.

However credit scoring is frequently used to predict the risk of a customer's defaulting of loan.

2.2.Scoring Methods

There are different historical statistical methods that have been used for calculating and developing credit scoring. Those are: linear probability models, logistic regression models, and probit model and discriminant analyses models. The first three use historical data for finding the probability of default. The discriminant analysis divides borrowers into high and low default risk classes. In this thesis we will be using a widely used method of credit scoring - logistic regression.

"Two newer methods beginning to be used in estimating default probabilities include options-pricing theory models and neural networks. These methods have the potential to be more useful in developing models for commercial loans, which tend to be more heterogeneous than consumer or mortgage loans, making the traditional statistical methods harder to apply." (Mester 1997)

The best methodology for credit scoring model has not been produced yet, since it depends on the dataset characteristics.

2.3.Credit Scoring History And Using Area

Credit scoring become widely used after 1980s.

"In the 1980s, the success of credit scoring in credit cards meant that banks started using scoring for other products, like personal loans, while in the last few years, scoring has been used for home loans and small business loan." (Thomas, Edelman and Crook 2002, p4)

But its history started much earlier, "when Sears used credit scoring to decide to whom to send its catalogues "(Lewis 1992). It was in the 1950s in America where credit lenders decided to make more accurate system to calculate scorecard.

In the past only banks used credit scoring, but then it was widely used for issuing credit cards, in other type of loan. Nowadays it is used in credit card, club card, mobile phone companies, insurance companies and government departments.

"The first Banks to use scoring for small-business loans were larger banks that had enough historical loan data to build a reliable model." (Mester 1997)
Credit scoring is likely to change the nature of small-business lending. It is not useful for large commercial loans.

2.4.For Lender And Customer

Credit scoring is useful from both lenders and customers' point of view.
Lenders. Credit scoring helps lender in the process of making decision to evaluate potential customers, to define their creditworthiness and avoid credit risk. By credit scoring lenders define who is worth to get credit or loan, at what interest rate and how much can be credit limit. Lenders can determine which customer would bring more gain. Also it takes less time and money in process.
Customers. Credit scores are one of the most important components of a consumer's personal finances. By controlling credit score customers can develop it and change the lenders result. It can save thousands of dollars depending how good the score is.
The more negative information is in the credit report, the lower credit score will be.
The credit scoring can help to avoid unnecessary credit risk.

2.5.Benefits Of Credit Scoring

There are three obvious benefits of credit scoring (Mester 1977)
"Quicker- when use with an automated software system, each customer is evaluated in second. For small-business it takes about 12-1/2 hours. Credit scoring can reduce this time.
Cheaper- this time savings means cost saving to the bank and benefits the customers as well. Customers need to provide only the information used in the

scoring system, so applications can be shorter, and scoring systems themselves are not prohibitively expensive.
More objective- objectivity helps lenders ensure they are applying the same underwriting criteria to all borrowers regardless of race, gender, or other factors in making credit decisions." (Mester 1997)

2.6 .Different Type Of Credit Scoring

There is number of several credit score formulas in use, each with different characteristics:

"The FICO score- This is the most widely adopted credit score and scoring model in the industry. The Fair Isaac Corporation is the father of the FICO score and is the originator of the credit report concept. The FICO score scale runs from 300 to 850 points." (About FICO Scores-see detail

(http://www.myfico.com/crediteducation/creditscores.aspx))

In fact, the FICO scores are not directly traded to customers. Instead, there are three main vendors - Experian, Trans Union, and Equifax, all using FICO score as a raw score, making some modification and selling the scores to loan applicants. All three maintain records of customers' credit history known as credit files. The customer's Credit Score is based on the information in your credit file at the time it is requested.

"The PLUS Score, with scores ranging from 330 to 830, is a user-friendly credit score model developed by Experian to help you see and understand how lenders view your credit worthiness. Higher scores represent a greater likelihood that you'll pay back your debts so you are viewed as being a lower credit risk to lenders. During the time your information can change, your credit score may be different from time to time.

(https://member.freecreditreport.com/scores/articles/different-types-of-scores)

"The Vantage Score- Vantage Score is a new credit scoring model created by America's three major credit reporting agencies to support a consistent and accurate

approach to credit scoring. This score provides lenders with nearly identical risk assessment across all three credit reporting companies. The Vantage Score scale ranges from 501 to 990." (Guina 2011)

No matter which scoring models lenders use, it pays to have great credit score. Customer's credit score affects whether he/she gets credit or not, and how high his interest rate will be. A better score can lower customer's interest rate.

FICO Scoring Method

Today in America the most widely used scoring method is still FICO. As it was mentioned already, its scale range is between 300 and 850. The highest score is obtained by a very small number of customers only. The vast majority of people will have scores between 600 and 800, as it is seen from the distribution given below. A score of 720 or higher will get you the most favorable interest rates on a mortgage, according to data from Fair Isaac Corporation.

By Fair Isaac Corp. reports, the American public's credit scores break out along these lines:

FICO credit score - Percentage
499 and below -2 percent
500-549 - 5 percent
550-599 - 8 percent
600-649 -12 percent
650-699 -15 percent
700-749 -18 percent
750-799 -27 percent
800 and above -13 percent

In determining the FICO score, mathematical models are used to analyze the data on an applicant's credit report. The FICO credit scoring formula is a closely guarded secret. However, it is known that it takes into consideration five main factors: previous credit performance, current level of indebtedness, time credit has

been in use, types of credit available and pursuit of new credit. More precisely, the FICO-scoring model looks at more than 20 specific factors in five categories.

1. Payment history (35%) -how you pay your bills.

The most important factor is how you have paid your bills in the past, placing the most emphasis on recent activity. Paying all your bills on time is good. Paying them late on a consistent basis is bad. Having accounts that were sent to collections is worse. Declaring bankruptcy is the worst.

2. Amount owned (30%) -amount of money you owe and the amount of available credit.

The second most important area is your outstanding debt -- how much money you owe on credit cards, car loans, mortgages, home equity lines, etc. Also considered is the total amount of credit you have available. If you have 10 credit cards that each have $10,000 credit limits, that's $100,000 of available credit. Statistically, people who have a lot of credit available tend to use it, which makes them a less attractive credit risk.

3. Length of credit history (15%)

The longer you've had credit, the more points you get.

4. Type of credit – mix of credit (10%)

The best scores will have a mix of both revolving credit, such as credit cards, and installment credit, such as mortgages and car loans. "Statistically, consumers with a richer variety of experiences are better credit risks," Watts says. "They know how to handle money." Many open accounts can have a negative impact, whether you are using the accounts or not.

5. New credit applications (10%)

How many credit applications you're filling out. Opening new accounts in a short period on time may negatively impact your score. (Myfico-the details can be obtained e.g. from

(http://www.myfico.com/crediteducation/whatsinyourscore.aspx))

3.Study Design And Data Collection

3.1. Target Population

Target population of this thesis are the individual who decided to apply for a credit first time, and individuals who already have credit history. All important information will be gathered from target applicants.

The identified pattern is used to predict the behavior of the future applicants based on the input or independent variables like income, job, debt etc. The same concept will be applied in this study, also for the default risk prediction of applicants.

Factors causing credit risk are very different, because of the variety of the borrowing populations. During gathering the information it is very important to consider all factors that are used for calculating credit scores, factors such as age, residual status, job type, income, etc.

3.2.Risk Factors

Scoring systems start using the best factors of variable to identify default risk.

By "(Jentzsch 2007), the general structure of credit scoring models is

$$S = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n ,$$

Where

S is dependent variable (measuring the customer's creditworthiness),

$x_1, \ldots, x_n$ are independent variables (customers attributes).

Note that the same is true for current thesis, because we will also use linear combinations of risk factors in our model.

Some attributes give information about the stability of the applicant (for example, time at present address, time at present employment), some characteristics present

information about the financial capacity of the applicant (for example income, credit cards, time with current bank, credit history), some variables provide information about the applicant's resources (for example, residential status, type of job, age).

Generally speaking, there are 5 C's of credit factors, each measured by specific variables, which are used more frequently in credit scoring. (The "Five Cs" of credit: (http://www.handsonbanking.org/financial-education/adults/the-five-cs-of-credit/).)

Character: Character is very important factor but not so easy to be measured. Expert must identify if applicant really repay loan or not. For this there are some factors that will give more clearly information about applicant: applicant's job, income stability, his/her credit history and personal characters.

Capacity: it is also very important factor. It includes information about applicant's financial capability, if he/she can repay loan. Income statement, dividends, also applicant expenses are considered in measurement.

Capital: It gathers information about applicants' assets, for example homes, boats, airplanes and etc. It will be considered if applicant wants to takes huge loan. Experts want to know if there are any additional backup capacities in case of unfavorable situation.

Collateral: If the situation occurs in which applicant is not able to repay by primary source, the mentioned factor will work as a secondary source of repayment. Collateral is the applicant asset what was put in financial institute instead of taken loan. Financial institutes can takes this collateral and sell it in the market.

Condition :This factor is depends on applicants job type and nature of the firms, where he/she works, it consider information about this firm's economic condition, that may have influence on applicants repayment of the loan.

3.3. Data Collection

Data were collected randomly from different banks and credit companies in Georgia. The sample consisted of 500 applicants. Data set is not very big and the reason of this is following: still a low number of people are looking for credit, because people think that getting credit is too risky.

Data such as applicant age, residual status, place of current residence, information about working, income, time at the same job, time with bank, all information about credit or loan, if applicant has defaulted on previous loan, if applicant missed any payment, how many times did applicant apply for credit or loan, what type of loan he/she has - all these details are important factors and must be collected by credit expert for the future credit decision.

The data consists of two groups: good customers (Y=1) who paid their loan back, and bad customers (Y=0) who defaulted on their loans. Each customer is described by 17 variables, shown in Annex 1. The variables describe consumer's profile and financial data.

Summary of Dataset

Number of applicants	500
Number of attributes	17

Among the 17 variables there are 16 independent (input) variables and 1 dependent (output) variable. Three (3) independent variables are "Scale" (numerical) variables and 13 variables are "Nominal" variables (but still with ordered categories). Each of nominal attributes has a scale ranging from 1 to K, with K depending on the attribute. Maximum sum of scores is 56. The output

variable Y is binary, taking values 0 and 1. All the information about variables is given in Annex 1 and Annex 2.

3.4. Data Preparation

After the data collection, data preparation is a very important part of the study. It is needed to identify incorrect information, to handle missing data in the dataset, etc. Only if the data is prepared properly and accurately, the model can give good result and credit expert will be able to make correct decisions.

4.Logistic Regression Analysis

4.1.General Form Of Logistic Regression Model

As we have already noted, logistic regression is one of the most frequently used statistical model used in credit scoring. It is the best to show probability of default and risk of decision.

Logistic regression models the relationship between a set of independent variables and the probability that a case is a member of one of the categories of the dependent variable. In our case, the two categories of the dependent variable Y are 1 (good customer) or 0 (bad customer). The frequencies of two categories in our data were 285 (good customers) and 214 (bad customers).

So, we decided to use logistic regression for the development of credit scoring. As we already said, logistic regression is a standard statistical technique for estimating the probability of default on loan performance and characteristics of the borrower based on historical data. Depending on the values of attributes (independent variables), we will find the probability that the dependent variable takes value 0 (default probability). For all necessary calculations, we use logistic regression procedure of the IBM SPSS statistical software.

By the logistic regression model the probabilities that $Y = 0$ (individual is bad) and $Y = 1$ (individual is good) are expressed as in equations (1) and (2):

$$P(Y = 0|X) = P = \frac{e^{\beta' X}}{1+e^{\beta' X}} \quad , \tag{1}$$

$$P(Y = 1|X) = 1 - P = \frac{1}{1+e^{\beta' X}} \quad . \tag{2}$$

The symbol β' stands for the vector of coefficients, $\beta' = (\beta_0, \beta_1, \dots, \beta_{16})$, and x is the column vector of independent variables, $x' = (x_0, x_1, \dots, x_{16})$.

The equations (1) and (2) are equivalent to (3):

$$\ln\left(\frac{P}{1-P}\right) = \beta' X =: l \tag{3}$$

Where *l* denotes the logit function of the probability *p*. Equation (3) is an indicator of linear relation between independent variables and logit function of the dependent variable. The coefficients β are estimated by the maximum likelihood method. We can write

$$P(Y = y_i) = P_i^{1-y_i}(1 - P_i)^{y_i} \quad (4)$$

Where the variable P_i is default probability in the i^{th} observation and y_i is the value of random variable *Y* that can be 0 or 1. Assuming that our *n* observations are independent, the likelihood of the data will thus be equal to

$$L = \Pi_{i=1}^{n} P_i^{1-y_i}(1 - P_i)^{y_i} . \quad (5)$$

According to maximum likelihood method, the function *L* is to be maximized over all possible values of the beta-coefficients. It can be completed by means of different software, including Excel (although it needs more work with Excel than with more specialized software). We have used IBM SPSS statistical package to estimate the parameters of logistic regression and to obtain other output information which helps to understand the quality of the model.

As we have noted already, there are 16 independent variables in the model. The logit variable *l* is the linear combination of the 16 independent variables weighted by the logistic coefficients:

$$l = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 \ldots\ldots + \beta_{16} * X_{16}.$$

Some modeling guidelines for logistic regression

There are some useful hints to be taken into account when applying logistic regression

(http://appricon.com/index.php/logistic-regression-analysis.html).

1) Data set should contain at least 30 rows of data. (This requirement is well satisfied in our case.)

2) The logistic regression model should be comprised of no more than 1 variable per 30-50 data rows. (This requirement is just fulfilled in our application, as 500/17 is close to 30).

3) A logistic regression model should have preselected variables used as the model core and defined by a professional in the field of application being modeled. The preselected variables are ones that considered as affecting the decision being modeled prior to the modeling process. The general approach is that a logistic regression model has to be based on the field of application and cannot be defined solely on statistical tests.
4) Logistic regression models should have a minimal set of variables. This rule cannot be quantified yet variables that add little to model performance should not be included. (We try to follow this rule by using special variable selection procedures like FORWARD).
5) The desired parameter values in the process of analysis are not absolute and relate to the field being modeled (example: models involving human behavior might have larger p-values and less accuracy compare to models involving physical phenomena).

4.2.Logistic Regression With All Covariates

In this work the IBM SPSS software was used to execute logistic regression.
We gathered information from 500 applicants, good and bad, and calculated credit scoring model by logistic regression procedure of IBM SPSS.
After processing the data in SPSS, the statistical outputs were offered. In the table "Case Processing Summary" it is seen that there are 499 cases (observed applicants) used, out of 500, in logistic regression. One case was not used for the reason that it contained missing data (erroneous non-numeric data).

Case Processing Summary			
Unweighted Cases		N	Percent
Selected Cases	Included in Analysis	499	99.8
	Missing Cases	1	.2
	Total	500	100.0
Unselected Cases		0	.0
Total		500	100.0

Let us see what's happened when we used all 16 independent variables as a predictors in modeling. After using SPSS logistic regression we obtained the table where each variable has its beta coefficient together with significance statistics.

Variables in the Equation								
Var[a]	B [b]	S.E. [c]	Wald [d]	Df [e]	Sig. [f]	Exp(B)[g]	95% C.I.for EXP(B)	
							Lower	Upper
V1	.049	.021	5.390	1	.020	1.050	1.008	1.095
V2	.048	.014	11.268	1	.001	1.050	1.020	1.080
V3	-.677	.259	6.806	1	.009	.508	.306	.845
V4	.317	.158	4.009	1	.045	1.373	1.007	1.872
V5	.627	.282	4.940	1	.026	1.871	1.077	3.252
V6	.116	.206	.319	1	.572	1.123	.750	1.682
V7	.673	.300	5.050	1	.025	1.961	1.090	3.527
V8	-.132	.240	.303	1	.582	.876	.548	1.402
V9	.298	.250	1.414	1	.234	1.347	.825	2.199
V10	-.226	.178	1.606	1	.205	.798	.562	1.132
V11	.768	.158	23.493	1	.000	2.156	1.580	2.941
V12	.469	.157	8.923	1	.003	1.599	1.175	2.175
V13	.000	.000	3.586	1	.058	1.000	1.000	1.000
V14	1.016	.203	25.076	1	.000	2.762	1.856	4.111

V15	.695	.201	11.947	1	.001	2.004	1.351	2.973
V16	.359	.176	4.155	1	.042	1.431	1.014	2.021
Constant	-15.850	1.929	67.501	1	.000	.000		

Explanations of column labels:

a.Variables included in the model: *V1, V2… V16.*

b.B (beta coefficients) gives information about linear relationship between independent and dependent variables, where the dependent variable is on the logit scale:

*l = log (odds) = -15.850+0.49*V1+…..+0.359*V16*

For the independent variables which are not significant, the coefficients are not significantly different from 0, which should be taken into account when interpreting the coefficients.

c.S.E. - These are the standard errors associated with the coefficients. The standard error is used for testing whether the parameter is significantly different from 0. The standard errors can also be used to form a confidence interval for the parameter.

d.Wald -This is the Wald chi-square test that tests the null hypothesis that the constant equals 0. This hypothesis is rejected because the p-value (listed in the column called "Sig.") is smaller than the critical p-value of .05 (or .01). Hence, we conclude that the constant is not 0. Usually, this finding is not of interest to researchers.

e.Df - This is the degrees of freedom for the Wald chi-square test .There is given 1 degree of freedom for each variable (predictor) in the model.

f.Sig. Is p-value of significance test of beta. Usually the p-value should be less than 0, 05 in order to include the variable into the model.

g.Exp(B) - These are the odds ratios for the predictors. They are the exponentiation of the beta-coefficients.

Looking at the p-values (located in the column labeled "Sig."), we find that there are 5 variables with significance higher than 0.05 (5%). (Therefore, we later repeat the whole analysis using stepwise variable selection procedure to exclude insignificant regressors).

Next we present some output tables of logistic regression where different aspects of the quality of the model are tested. We give some explanations how to interpret these output tables of logistic regression (the details can be obtained e.g. from (http://appricon.com/index.php/logistic-regression-analysis.html).)

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	417.466[a]	.411	.552

In this summary we can see that -2 log likelihood statistic is 417.466. If statistic was smaller, the model would be better.

Cox and Snell R^2 test as well as other logistic regression R^2 tests tries to measure the strength of association of the model. The values of this test are between 0 and 1. The Nagelkerke R^2 (a modification of the Cox and Snell R^2) is more common and considered a better indication to strength of association. Our R2=0.552 is not very high, so this relationship is not very strong.

Hosmer-Lemeshow table is a model classification table which describes both expected model classifications and actual model classifications. The Hosmer-Lemeshow table divides the data into 10 groups (declines, one per row) each representing the expected and observed frequency of both 1 and 0 values. The expected frequency of data assigned to each declines should match the actual frequency outcome and each declines should contain data.

Partition for the Hosmer and Lemeshow Test

Group	Total	Y = 1		Y = 0	
		Observed	Expected	Observed	Expected
1	50	6	1.88	44	48.12
2	50	0	7.03	50	42.97
3	50	13	14.18	37	35.82
4	50	17	20.11	33	29.89
5	50	29	26.65	21	23.35
6	50	35	33.96	15	16.04
7	50	39	40.54	11	9.46
8	55	55	49.11	0	5.89
9	50	47	47.67	3	2.33
10	44	44	43.89	0	0.11

We can see that the differences between observed and expected frequencies are not big with our model (sometimes they are very close).

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	26.302	8	.001

Hosmer-Lemeshow Probability Test is based on a chi-square test which is done over the Hosmer – Lemeshow table (above). This important parameter tests the assumption that the model distinguishes the explained variable better. The actual Null hypothesis is that the model is insignificant and the test tries to break this hypothesis. Values for this test should be higher than 0.5 – 0.6. It is .001 in our case which means that the differences between observed and predicted

frequencies cannot be explained by chance only - it is also the problem of model inadequacy.

Classification tables. In binomial logistic regression, the classification table is a table that contains the observed and predicted model results. Each data record is classified using the computed probability given by the model (a value between 0 and 1) and the cut value which is the minimal value of probability that should be classified as 1. The default "cut value" value is 0.5, determines that a data record that has a value larger than 0.5 should be classified as 1. In our analysis, we have used several other cut probabilities as well: 0.3, 0.4, 0.5, 0.55, 0.6, and 0.7.

Classification Table									
	Correct		**Incorrect**		**Percentages**				
Prob Level	**Event**	**Non-Event**	**Event**	**Non-Event**	**Correct**	**Sensi-tivity**	**Speci-ficity**	**False POS**	**False NEG**
0.300	267	122	92	18	78.0	93.7	57.0	25.6	12.9
0.400	253	136	78	32	78.0	88.8	63.6	23.6	19.0
0.500	236	165	49	49	80.4	82.8	77.1	17.2	22.9
0.550	228	172	42	57	80.2	80.0	80.4	15.6	24.9
0.600	220	178	36	65	79.8	77.2	83.2	14.1	26.7
0.700	195	190	24	90	77.2	68.4	88.8	11.0	32.1

Each row of the classification table corresponds to a specific cut probability and has 4 data cells:

1. Observed 0 Predicted 0 – The number of cases that were both predicted and observed as 0. The model classification was correct for these records.
2. Observed 0 Predicted 1 – The number of cases that were predicted as 1 yet observed as 0. The records in this cell are referred to as **false negatives**. The model classification was incorrect for these records.
3. Observed 1 Predicted 1 – The number of cases that were both predicted and observed as 1. The model classification was correct for these records.

4. Observed 1 Predicted 0 – The number of cases that were predicted as 0 yet observed as 1. The records in this cell are referred to as **false positives** the model classification was incorrect for these records.

Different fields of applications require different rates of false positives and false negatives since in some applications false positives cannot be tolerated while in other applications, false negatives cannot be tolerated.

We next present the row corresponding to cut value 0.5 in a form of 2x2 classification table.

Classification Table (cut value 0.5)[a]

Observed		Predicted		
		Y		
		0	1	Percentage Correct
Y	0	165	49	77.1
	1	49	236	82.8
Overall Percentage				80.4

Classification Table shows accuracy of the model. This rule allow us to classify 236/285=82.8%, that this percentage of occurrences is correctly predicted, it is a sensitivity of prediction, P (correct/ event did occur). As we see 165/214=77.1% of the subject where the predicted event was not observed, P (correct/ event did not occur). As we see prediction 401 out of 500 times were correct, for an overall success rate of 80.4%.

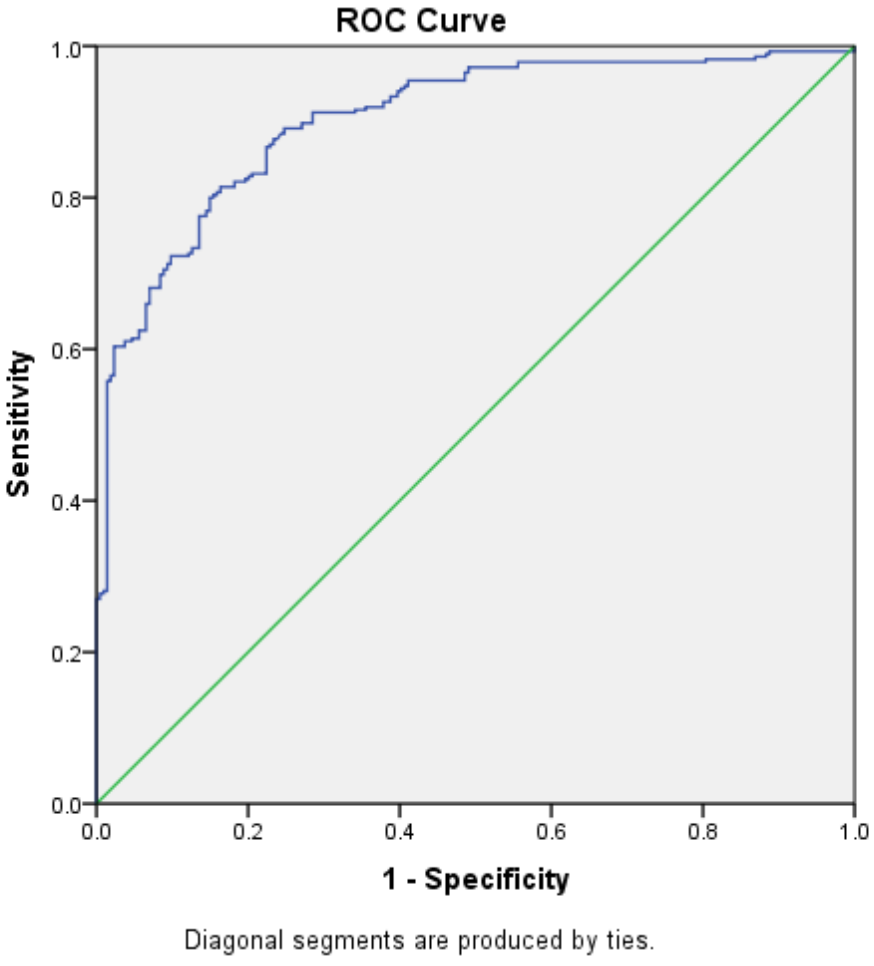
To finish, let us add a warning remark on how to use classification tables [17].

"Classification tables (the tables that show how model classified, the rate of hits/misses) should be assessed with caution. In order to achieve good and long lasting results, statistical testing should be the main tool of analysis and classification tables should be treated as an independent test conducted after model quality assessments are completed. Classification tables play an important

role once a model is deployed and used since only reality shows the true quality of the model (after deployment). At this stage classification tables computed over the model results are the main tool for logistic regression model performance analysis."

Roc curve

We calculated sensitivity (true positive) and specificity (true negative) pairs for all possible cutoff points from 0 to 1, and plot sensitivity on the Y axis and (1-specificity) on the X axis. This curve is called the receiver operating characteristic (ROC) curve. The area under the ROC curve ranges from 0.5 and 1.0 with larger values indicative of better fit.



Area under ROC Curve (AOC) is a good indication to model performance (values are between 0.5 and 1). This variable should be as high as possible with some restrictions. Typical values indicate the following:

- 0.5 – No distinguish ability (the model has no meaning).
- 0.51 – 0.7 – Low distinguish ability (not a very good model yet the model can be used).
- 0.71 – 0.9 – Very good distinguish ability.
- 0.91 – 1 – Excellent distinguish ability.

In some fields, logistic regression models can have an excellent distinguish ability, however this might indicate that the model is "too good to be true". One should double and triple check the model making sure that no variables from the future are present and that the model has no other odd parameter values.

In our case, the area under the curve is .903 with 95% confidence interval (.876, 929) (see the next table). Also, the area under the curve is significantly different from 0.5 since p-value is .000 meaning that the logistic regression classifies the group significantly better than by chance.

Area Under the Curve

Test Result Variable(s): Predicted probability

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.903	.013	.000	.876	.929

To conclude, the logistic regression analysis with all varieties (full model) produced a model with relatively high prediction ability. However, since it contains several variables with insignificant betas, we cannot be sure about the stability of the model. Therefore, we will proceed with logistic regression that uses a variable selection method to avoid insignificant and possibly useless variables in the model.

4.3. Logistic Regression With Selected Covariates

Here we apply logistic regression procedure FORWARD which automatically includes into the model only the varieties that are significant. A slightly simplified description of the procedure is the following.

1. Choose the variable with the smallest significance (p-value) of beta. If that significance is less than the probability for a variable to enter (0,05 in our case), then go to step 2.

2. Inclusion step: Update the current model by adding a new variable.

3. Exclusion step: If the largest significance in the current model is larger than the probability for variable removal (0,20 in our case), then remove respective variable from the model.

4. Based on the MLEs of the current model, calculate the score statistic for every variable eligible for inclusion and find its significance. If the smallest significance is less than the probability for a variable to enter, then go to step 2.

We applied the logistic regression FORWARD procedure to the same data as before. Application of the FORWARD procedure resulted in a model with 8 covariates, as seen in the next table.

Variables in the Equation

	B [b]	S.E [c]	Wald [d]	Df [e]	Sig. [d]	Exp(B) [f]	95% C.I.for EXP(B)	
							Lower	Upper
V1	.056	.013	19.132	1	.000	1.058	1.031	1.084
V2	.056	.011	27.838	1	.000	1.058	1.036	1.080
V4	.369	.144	6.533	1	.011	1.447	1.090	1.920
V5	.904	.184	24.024	1	.000	2.470	1.721	3.546

	V11	.761	.124	37.493	1	.000	2.140	1.678	2.731
	V13	.000	.000	2.517	1	.113	1.000	1.000	1.000
	V14	.759	.166	20.833	1	.000	2.137	1.542	2.961
	V16	.605	.157	14.809	1	.000	1.832	1.346	2.494
	Constant	-13.094	1.589	67.910	1	.000	.000		

The regressor variables included into the model are:

V1 Applicant age,

V2 Duration of credit in months,

V4 Living current place

V5 Type of job

V11 Last miss of payment

V13 Amount of credit

V14 Further debtors/guarantors

V16 How many times have you applied for credit in the past year

All the beta-coefficients are positive by sign. The beta for V13 is very small but it comes from diferent measurement scale (credit amount is measured in currency units, other variables mostly on 5-points scale).

There is one variable (V13) with large p-value 0.113 but this is permitted by the procedure, since this p-value is still less than the exclusion probability level 0.2.

The model summary is as follows.

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	441.337 [a]	.382	.513

In this summary we can see that -2log likelihood statistic is 441.337 which is higher than with full model where it was 417. Also the value of R2=0.513 shows

weaker predictive capacity of the model than before. The same says the Hosmer –Lemeshow test with its p-value 0.000 (before it was 0.001).

Hosmer and Lemeshow Test			
Step	Chi-square	df	Sig.
1	39.362	8	.000

We calculated predicted creditworthiness for each data point, and the residuals. By their actual value of Y, we have 285 good applicants and 214 bad applicants. The classification table below **shows** accuracy of the model. This rule allow us to classify correctly 243/285=85.3% of good applicants (sensitivity of the prediction rule= P (correct/event did occur). At the same time 172/214=80.4% of bad subjects were correctly classified (P(correct/ event did not occur)). As we see, in total 415 cases out of 499 were classified correctly, which makes the overall success rate be equal to 415/499=83.2%.

Classification Table[a]					
	Observed		Predicted		
			Y		Percentage Correct
			0	1	
Step 1	Y	0	172	42	80.4
		1	42	243	85.3
	Overall Percentage				83.2
a. The cut value is .500					

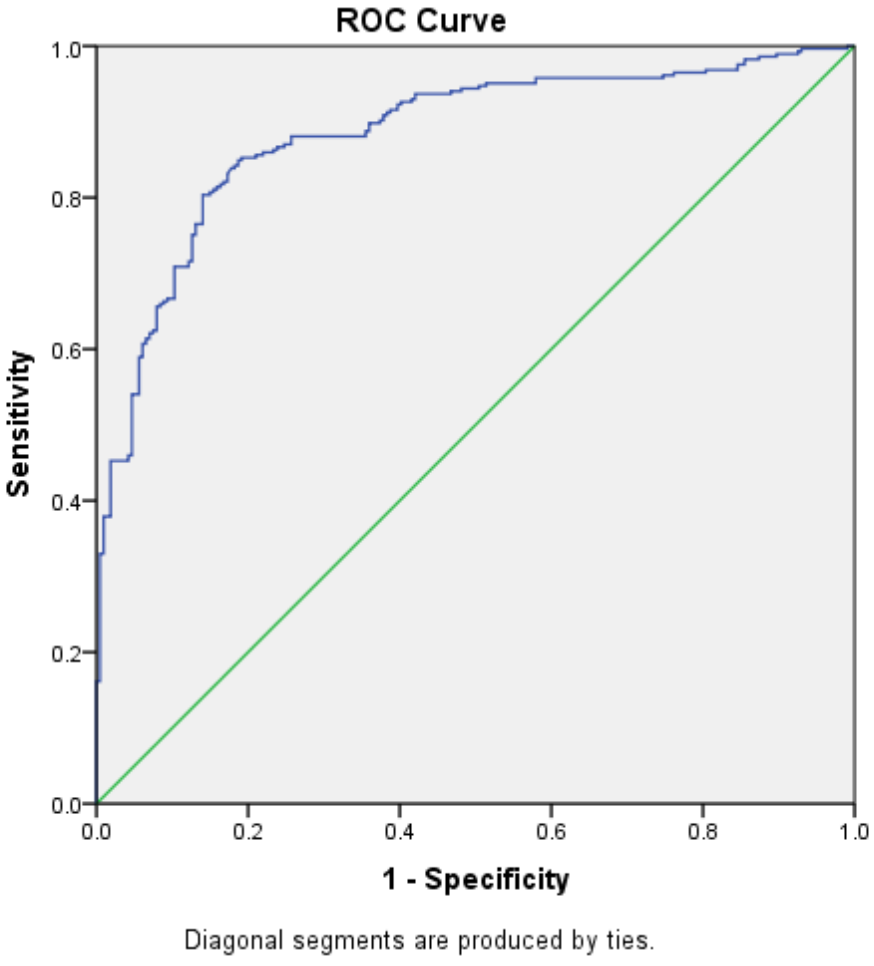
In classification table overall percentage gives information that in my model 83.2% of cases were correctly predicted. As I saw this percentage has increased from 80, 4% in case of the full model to 83.2% for selected variables model. Still, this does not say too much because both rules use same cutting value 0.5 but these

cutting values are not necessarily optimal for these two analyses. We can only say that in the case of the new model the cutting value 0.5 works well.

In annex 3 predicted default probabilities for bad applicants (Y=0) are given. We see that these probabilities are mostly large, as expected.
Similarly, in annex 4, predicted default probabilities for good applicants (Y=1) are given. We see that these probabilities are mostly small, as expected.

ROC curve

This time the ROC curve looks like follows:



The area under the curve is now .884 with 95% confidence interval (.854, .914) which is slightly less than in case of the full model.

Area Under the Curve				
Test Result Variable(s): Predicted probability				
Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.884	.015	.000	.854	.914

In general, one can see that the new model with only 8 covariates has almost the same quality as the full model with 16 variables. However, it is cheaper (less variables to measure) and more stable.

Logistic regression helps to make more accurate decision about credit accept and reject. By this statistical method lenders can easily avoid some risk that bring them loss and on the other side, to get new customers who bring more revenue.

Final Comment

Let us finish with a comment made by credit professionals about the role of statistical models ion credit scoring.

“Although credit risk assessment is one of the most successful applications of applied statistics, the best statistical models don’t promise credit scoring success, it depends on the experienced risk management practices, the way models are developed and applied, and proper use of the management information systems” (Mays 1998). “And at the same time, selections of the independent variables are very important in the model development phase because they determine the attributes that decide the value of the credit score, and the values of the independent variables are normally collected from the application form. It is very significant to identify which variables will be selected and included in the final scoring model. “

Conclusions

In this Master's thesis we used real data collected from credit customers, both "good" (with no problems in paying back the loan) and "bad" (defaulted). We had 16 scaled independent variables (covariates) that had influence of applicants' credit scoring in observed situation. We created a statistical model – logistic regression model - which calculates predicted probability of the default.

In creating the model, we used logistic regression FORWARD procedure in IBM SPSS. The model obtained consists only of 8 variables which are used in calculation of predicted default probabilities. With this model we found out that only 42 good and 42 bad applicants (less than 20 percent in both cases) were incorrectly classified. We also calculated the ROC curve of the model and found out that it is possible to find a cutting value such that both sensitivity and specificity are as high as 80%.

To conclude, logistic regression helps to make more accurate decision about credit accept and reject. By this statistical method lenders can easily avoid some risk that bring them loss and on the other side, to get new customers who bring more revenue.

.

References

1. About fico score
2. http://www.myfico.com/crediteducation/creditscores.aspx (last seen 10.04.2015)
3. Berjisian Adel, Sepehrdoust Hamid, The Application of Logistic Regression Analysis to Bank's Risk Management, 4-7pg.
4. Different types of credit score
5. https://member.freecreditreport.com/scores/articles/different-types-of-scores, about plus score. (last seen 10.04.2015)
6. Gouvêa Maria Aparecida, Gonçalves Eric Bacconi, (2007) Credit Risk Analysis Applying Logistic Regression, Neural Networks and GeneticAlgorithms Models.
7. Greene H. William, (1992) A Statistical Model for Credit Scoring.
8. Guina Ryan, (2011) examining different types of credit scoring,on pg The military wallet
9. Islam Samsul , Lin Zhou, Fei Li, (2009) Application of Artificial Intelligence (Artificial Neural Network) to Assess Credit Risk: A Predictive Model For Credit Card Scoring.
10. JENTZSCH, N. (2007) Financial Privacy: An International Comparison of Credit Reporting Systems (Contributions to Economics), Springer.
11. MAYS, E. (1998) Credit Risk Modeling: Design and Application, CRC.
12. MAYS, E. (2001) Handbook of Credit Scoring, Global Professional Publishing.
13. Mester J. Loretta ,(1997) ''what's the point of credit scoring?"
14. Thomas C. Lyn, Edelman B. David, and Cook N. Jonathan (2002) Credit scoring and its application (America, 2002) p.1 chp. 1.1.
15. Thomas C. Lyn, Edelman B. David, and Cook N. Jonathan (2002) Credit scoring and its application (America, 2002) p.4 chp 1.3.
16. What's my score http://www.myfico.com/crediteducation/whatsinyourscore.aspx, How my fico score is calculated (Seen 10.05.2015)

17. Mirta Bensic, Natasa Sarlija , Marijana Zekic-Susac: Modeling Small Business Credit Scoring by Using Logistic Regression, Neural Networks, and Decision Trees
18. Annotated SPSS Output Logistic Regression
http://www.ats.ucla.edu/stat/spss/output/logistic.htm
(Seen 10.05.2015)
19. http://appricon.com/index.php/logistic-regression-analysis.html
20. http://www.handsonbanking.org/financial-education/adults/the-five-cs-of-credit/
21. http://www.investopedia.com/terms/f/five-c-credit.asp

Annex 1: Dataset variables description

#	Variable	Description	Measure
1	Creditworthiness	Status of credit applicant	Nominal
2	V1	Applicant age	Scale
3	V2	Duration of credit in months	Scale
4	V3	Home status	Nominal
5	V4	Living current place	Nominal
6	V5	Type of job	Nominal
7	V6	Working current place	Nominal
8	V7	Monthly income	Nominal
9	V8	Time with bank	Nominal
10	V9	Available assets	Nominal
11	V10	Number of credit at the bank(including running one)	Nominal
12	V11	Last miss of payment	Nominal
13	V12	How long ago most negative event occurred	Nominal
14	V13	Amount of credit (sum of credit)	Scale
15	V14	Further debtors/guarantors	Nominal
16	V15	Low long ago was first credit	Nominal
17	V16	How many times have you applied for credit in the past year	Nominal

Annex 2: Variables, categories, scores

#	Variable	Categories/component	Score
Y	Creditworthiness	Good	1
		Bad	0
V1	Applicant age		
V2	Duration of credit in months		
V3	Home status	owner	3
		rent	2
		other	1
V4	Living current place	<1 year	1
		1<=…<4 years	2
		4<=…<7 years	3
		>=7 years	4
V5	Type of job	unemployed	1
		unskilled with permanent residence	2
		skilled worker/skilled employee	3
		self-employed	4
V6	Working current place	unemployed	1
		<=1 year	2
		1<=..<4 years	3
		4<=…<7 years	4
		>= 7 years	5
V7	Monthly income	None	1
		<=500 lari	2
		500<=…<1500 lari	3
		1500<=…<5000 lari	4

		>=5000 lari	5
V8	Time with bank	<=2 year	1
		2<=…<5 years	2
		5<=…<8 years	3
		>=8 years	4
V9	Available assets	Ownership of house or land	3
		Car/other	2
		no assets	1
V10	Number of credit at the bank(including running one)	None	1
		1 time	2
		2 or 3 times	3
		4 or 5 times	4
		>=6 times	5
V11	Last miss of payment	never	5
		6 month ago	1
		6<=…<12 month ago	2
		1<=…<2years ago	3
		>=2 years ago	4
V12	How long ago most negative event occurred	never	5
		<1 year ago	1
		1<=…<5 years ago	2
		5<=… <8 years ago	3
		>=8 years ago	4
V13	Amount of credit		
V14	Further debtors/guarantors	none	1
		Co-applicant	2
		Guarantor	3
V15		not yet	1

	How long ago was first credit	<1year	2
		1<=..<3 years ago	3
		3<=…< 5 years ago	4
		>=5 years ago	5
V16	How many times have you applied for credit in the past year	None	3
		1 time	5
		2 or 3 times	4
		4 or 5 times	2
		>=5 times	1

Annex 3: Actual credit, predicted probability

Y	Predicted deafault probability
0	98%
0	94%
0	92%
0	91%
0	87%
0	87%
0	87%
0	86%
0	86%
0	85%
0	80%
0	79%
0	75%
0	74%
0	72%
0	70%
0	69%
0	69%
0	68%
0	66%
0	65%
0	64%
0	62%
0	61%

Annex 4: Actual, predicted

Y	Predicted default probability
1	2%
1	2%
1	2%
1	3%
1	6%
1	7%
1	7%
1	9%
1	9%
1	15%
1	15%
1	17%
1	26%
1	26%
1	28%
1	29%
1	31%
1	32%
1	34%
1	34%
1	35%
1	35%
1	37%
1	37%
1	38%

Non-exclusive licence to reproduce thesis and make thesis public

I, Salome Tabagari
(*author's name*)
(date of birth: 02.11.1988),

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 web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

Credit Scoring by Logistic Regression,
(title of thesis)

supervised by Prof. Kalev Pärna,
(supervisor's name)

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

3. I 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, **dd.mm.yyyy** 13.05.2015