University of Tartu

Faculty of Science and Technology

Institute of Mathematics and Statistics

Tevfik Can Özay

Credit Scoring by Segmented Modelling

Financial Mathematics

Master Thesis (15 ECTS)

Supervisor: Kalev Pärna

Tartu 2016

Credit Scoring by Segmented Modelling

Master's thesis

Tevfik Can Özay

Abstract. This study is devoted to small loan evaluation modelling which is known as credit scoring. Credit scoring models help the decision takers (such as credit offices, banks …) decide customers' creditworthiness in short time without prejudice. Main goal of this master thesis was to understand feasibility and effectiveness of credit scoring model by using logistic regression technique and obtaining important variables for credits scoring models. Furthermore we targeted to reveal how segmentation (creating different score cards for different age groups) can help to predict more accurately. In this study, we worked with real data which was provided by local company in Estonia. In conclusion, our results showed that credit scoring by logistic regression helped to discriminate good customers effectively and the use of segmentation improves the model's accuracy.

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

Key words: Credit Scoring, Logistic Regression, Regression Analysis

Krediidiskooring segmenteeritud mudelite abil

Magistritöö

Tevfik Can Özay

Lühikokkuvõte. Antud töö on pühendatud väikelaenude hindamise modelleerimisele, mis on tuntud krediidiskooringu nime all. Krediidiskooring võimaldab laenuandjatel (laenuasutused, pangad) otsustada erapooletult laenutaotleja krediidikõlblikkuse üle. Antud töö peaeesmärgiks oli välja selgitada logistilise regressiooni kasutamisvõimalused krediidiskooringu mudeli loomisel ning tuvastada sellise mudeli tähtsad argumenttunnused. Lisaks on püütud välja selgitada segmenteerimise osa mudeli prognoosivõime parandamisel, luues erinevatele vanuserühmadele erinevad mudelid. Töö empiirilises osas on kasutatud reaalseid andmeid. Kokkuvõttes näitavad töö tulemused, et logistiline regression võimaldab efektiivselt eristada häid kliente halbadest ning et segmenteerimine aitab parandada mudeli täpsust.

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

Võtmesõnad: krediidiskooring, regressioonanalüüs, logistiline regressioon

Table of contents

Introduction

Nowadays credit sector is continuing to grow rapidly and in the meantime importance of credit scoring and managing the process of credit assessments is getting more critical for lenders, credit offices and banks. Emerging issue is to handle dramatically increasing number of applications whether candidates are eligible to take credit or not and having limited time to evaluate all these applications. Most of the professional lenders are already using statistical modelling techniques to assess their potential candidates. On the other hand, traditional judgmental assessment is still in common usage even though it is quite long process and it is needed massive experience and trainings. Therefore credit scoring has very important role which provides opportunity to assess candidates in short period and without prejudice. In this point, there are so many doubts about which statistical methods are more helpful and feasible for different credits such as small loans, mortgage … And there are many questions about what is benefit of creating different scorecards/models for different groups? Is creating different scorecards worth to work on it? How effectively does it help to improve accuracy of predictions? Increased competition and growing pressures for revenue generation have led credit-granting and other financial institutions to search for more effective ways to attract new creditworthy customers and at the same time, control losses [6]. Therefore professionals are always trying to find better and more accurate method for credits scoring in order to have better predictions. In this study, we will work with logistic regression which is one of the most popular methods for creating credit scoring model recently and analysis is performed by IBM SPSS 21.

Main purpose of this master thesis is 1) to show effectiveness of credit scoring model by logistic regression method; 2) to detect significant parameters for credit scoring model which may help further researches; 3) to understand importance of segmentation by comparing segmented models with one single model for everyone. With this purposes, we consider classification table as main indicator in this study which is one of the ways to see the percentage of correct prediction of each model.

The first chapter focuses on what is credit scoring and gives an historical background. It is followed by logistic regression in Chapter 2. The next chapter deals with framework of empirical study and data preparation. We emphasize our target population, dependent and independent variables, detecting outliers and how we treat missing values in this study. In Chapter 4, we concentrate on creating one single model for all clients by using logistic regression and in Chapter 5, we create 4 different models for the same number of age groups.

In chapter 6, there is summary of our results and comparison of two different approaches in order to see benefits from creating different score cards for different age groups.

1 Credit Scoring

1.1. What is Credit Scoring?

In current context, ‘credit’ simply means, “buy now, pay later”, whether the purchase is for short-term consumption, durable goods and other assets that provide users with valuable services or productive enterprises. The word ‘credit’ comes from the old Latin word ‘credo’, which means, ‘trust in’ or ‘rely on’. If you lend something to somebody, then you have to have trust in him or her to honor the obligation. [1]

Credit Scoring is set of decisions model and their underlying techniques that aid lenders in the granting of consumer credit. These techniques decide who will get credit, how much credit they should get, and what operational strategies will enhance the profitability of borrowers to the lenders. [8] According to Anderson, Credit scoring is the use of statistical models to transform relevant data into numerical measures that guide credit decisions.

1.2 History of Credit Scoring

Following overview is based on [8]

Mostly it is believed that people started borrowing and repaying while human started to communicate with each other. The first recorded instance of credit comes from ancient Babylon. According to stone table around 2000 BC, two shekels of silver have been borrowed by Mas-Schamach, son of Adadrimeni, from the sun priestess Amat-Schamach, the daughter of Warad-Enlil. He will pay the Sun-God’s interest. At the time of harvest he will pay back the sum and interest upon it.

The next thousands years, the “Dark Ages” of European history, saw a little development in credit, but by the time of the Crusades in the thirteen century, pawn shop has been developed however Crusades had no interest in this time. At the same time, merchants quickly saw the possibilities and by 1350 commercial pawn shops charging interest were found throughout Europe. During the middle ages, there was ongoing debate on the morality of charging interest on loans. The outcome of the debate in Europe was that if the lender levied small charges, this was interest and was acceptable, but large charges were usury, which was bad. Even Shakespeare got into this debate with his portrait of the Merchant of Venice. Also at this time, kings and potentates began to have to borrow in order to finance their wars and other expenses.

Lending was more politics than business. The rise of the middle classes in the 1800's led to the formation of a number of private banks, which were willing to give bank overdrafts to fund businesses and living expenses. However, this start of consumer credit was restricted to a very small proportion of the population.

The real revolution started in the 1920's when consumers started to buy motor cars. Finance companies were developed to respond to this need and experienced rapid growth before World War II. At the same time, mail-order companies began to grow as consumers in smaller towns demanded the clothes and household items that were available only in larger population centers. These were advertised in catalogues, and the companies were willing to send the goods on credit and allow customers to pay over an extended period.

While the history of credit stretches back 5000 years, the history of credit scoring is only 50 years old. Credit scoring is essentially a way to identify different groups in a population when one cannot see the characteristic that defines the groups.

During the 1930's some mail-order companies had introduced numerical scoring systems to try to overcome the inconsistencies in credit decisions across credit analysts. With the start of the World War II, all the finance houses and mail-order firms began to experience difficulties with credit management. Hence the firms had the analysts write down the rules of thumb they used to decide to whom to give loans. These rules were then used by non-experts to help make credit decisions.

It did not take long after the war ended for some folks to connect the automation of credit decisions and the classification techniques being developed in statistics and to see the benefit of using statistically derived models in lending decisions. The first consultancy was formed in San Francisco by Bill Fair and Earl Isaac in the early 1950's, and their clients were mainly fiancé houses, retailers and mail-order firms. The arrival of credit cards in the late 1960's made the banks and other credit card issuers realize the usefulness of credit scoring. The number of people applying for credit cards each day made it impossible in both economics and manpower terms to do anything but automate the lending decision. The growth in computing power made this possible. The organizations found credit scoring to be a much better predictor than any judgmental scheme and default rates dropped by 50 percent or more. In the 1980's, 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 loans. In the 1990's, growth in direct marketing led to the use of scorecards to improve the response rate to advertising campaigns. Advances in computing allowed other techniques to be tried to build scorecards. In the 1980's, logistic regression and linear programming, the two main stalwarts of today's card builder, were introduced.

1.3 Credit Assessment Before Credit Scoring and Why Credit Scoring

This section is based on [8] and [5]

In early 1960's, lenders had some difficulties to assess their customers individual applications. Traditional assessment needed so long time for evaluating only one candidate. Traditional assessment simply relied on evaluating borrowers' characteristic, ability to pay back his or her debt and individual experience of investigator. Generally manager took responsibility to assess their clients. Investigator evaluated the length of relationship with customer, likelihood of payment, assess stability, honesty and other characteristics. If manager did not convince and had doubts about customers, it meant further investigations and spent more time for single candidate. Process was obviously slow and inconsistent. There were plenty of disadvantages for processing traditional assessment;

- Task manager needed education and so long training period,
- Process was extremely slow and it was big obstacle to prevent having more customers/lenders,
- There was no unbiased assessment opportunities,
- Borrower needed to have quite long history to be invited investigation appointment.

Rapidly growing credit sector needed new assessment system. And many changes occurred in lender and borrower environment. Some of these changes were as follows;

- Banks changed their market position considerably and began to market their products. This in turn meant that they had to sell products to customers not only whom they hardly knew but whom they had enticed.
- There was phenomenal growth in credit cards. Sales authorizations of this products meant that there had to be a mechanism for making a lending decision very quickly and around the clock. Also, volumes of applications were such that the manager or other trained credit analyst could would not have the time or opportunity to interview all the applicants. Clearly there would be insufficient numbers of experienced bank managers to handle the volume. During the 1980's, a handful of U.K. operations were dealing with several thousand applications each day.

- Banking practice changed emphasis. Previously, banks had focused almost exclusively on large lending and corporate customers. Now consumer lending was an important and growing part of the bank. It would still be a minority part by value but was becoming significant. Bank could not control the quality across a branch network of hundreds or thousands of braches and mistakes were made. With corporate lending, the aim was usually to avoid any losses. However, banks began to realize that with consumer lending, the aim should not be to avoid any losses but to maximize profits. Keeping losses under control is part of that, but one could maximize profits by taking on a small controlled level of bad debts and so expand the consumer lending book.

1.4 Statistical Methods For Building Credits Scoring

When credit scoring was first developed in the 1950s and 1960s, the only methods used were statistical discrimination and classification methods. Even today statistical methods are by far the most common methods for building credit scorecards. Statistical techniques allow one to identify and remove unimportant characteristics and ensure that all the important characteristics remain in model.

Although statistical methods were the first to be used to build scoring system and they remained the most important methods, there have been changes in the methods used during the intervening 40 years. Initially, the methods were based around the discrimination methods suggested by Fisher (1936) for general classification problems. This led to a linear scorecard based on the Fisher linear discriminant function. The assumptions that were needed to ensure that this was the best way to discriminate between good and bad potential customers were extremely restrictive and clearly did not hold in practice, although the scorecards produced were very robust. The Fisher approach could be viewed as a form of linear regression, and this led to an investigation of other forms of regression. By far the most successful of these is Logistic Regression, which has taken over from the linear regression-discriminant analysis approach as the most common statistical method. Another approach that has found favor over the last 20 years is the classification tree approach. With this, one splits the set of applicants into a number of different subgroups depending on their attributes and then classifies each subgroup as satisfactory or unsatisfactory. Although this does not give a weight to each of the attributes as the linear scorecards does. There are also few non parametric approach based on nearest neighbors. [8]

2 Logistic Regression

This section is based on [9] and [6]

In the classical regression framework, we are interested in modeling a continuous response variable y as a function of one or more predictor variables. Most regression problems are of this type. However, there are numerous examples where the response of interest is not continuous, but binary. Consider an experiment where the measured outcome of interest is either a success or failure, which we can code as a 1 or a 0. The probability of a success or failure may depend on a set of predictor variables. One idea on how to model such data is to simply fit a regression with the goal of estimating the probability of success given some values of the predictor. However, this approach will not work because probabilities are constrained to fall between 0 and 1. In the classic regression setup with a continuous response, the predicted values can range over all real numbers. Therefore, a different modelling technique is needed. That is, in with a binary outcome, regression of y on x is a conditional probability. If we label $y = 1$ as a "success", then the goal is to model the probability of success given x. The approach to this problem illustrated here is known as Logistic Regression.

Logistic regression, like most other predictive modeling methods, uses a set of predictor characteristics to predict the likelihood (or probability) of a specific outcome (the target). The equation for the logit transformation of a probability of an event is shown by:

$$logit(\pi) = ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

π = posterior probability of "event", given inputs,

x = input variables,

β_0 = intercept of regression line,

β_i = parameters

Logistic Regression function for probability of individual is good:

$$\pi = \left(\frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k}}\right)$$

Logistic Regression function for probability of individual is bad:

$$(1-\pi) = \left(1 - \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}}\right)$$

Regression can be run to find out the best possible model using all options available. This is commonly known as "all possible" regression techniques and is computationally intensive, especially if there are a lot of independent input characteristics. Far more commonly used are the three types of stepwise logistic regression techniques:

Forward Selection: First selects the best one characteristic model based on the individual predictive power of each characteristic, then adds further characteristics to this model to create the best two, three, four, and so on characteristic models incrementally, until no remaining characteristics have *p*-values of less than some significant level , or univariate Chi Square above a determined level. This method is efficient, but can be weak if there are too many characteristics or high correlation.

Backward Elimination: The opposite of forward selection, this method starts with all the characteristics in the model and sequentially eliminates characteristics that are considered the least significant, given the other characteristics in the model, until all the remaining characteristics have a *p*-value below a significant level or based on some other measure of multivariate significance. This method allows variables of lower significance a higher chance to enter the model, much more than forward or stepwise, whereby one or two powerful variables can dominate.

Stepwise: A combination of the above two techniques, this involves adding and removing characteristics dynamically from the scorecard in each step until the best combination is reached. A user can set minimum p-values required to be added to the model, or to be kept in the model.

2.1 Why Logistic Regression Analysis for Credit Scoring

The Linear regression model provides a powerful device for organizing data analysis. Researchers focus on the explanation of dependent variable, Y, as a function of multiple independent variables, from X_1 to X_k . Models are specified, variables are measured and equations are estimated with ordinary least squares. All goes well if the classical linear regression assumptions are met. However several assumptions are likely to be unmet if the dependent variable has only two or three response categories. In particular, with a dichotomous dependent variable, assumptions of homoscedasticity, linearity and normality are violated and OLS estimates are inefficient at best. The maximum likelihood estimation of a logistic regression overcomes inefficiency, transforming Y $(1{,}0)$ into a logit. [4]

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. [3] When we consider Credit Scoring models, our main purpose is understanding difference between being Good Customer and Bad Customer. Status of customer have two options in this case. It means binary dependent variable where Logistic Regression provides best approach to deal with categorical dependent variable.

3 Framework of Study and Data Preparation

3.1 Target population

There are two most common approaches to create the credit scoring model. One is based on all component of population; Accepted bad and good clients and rejected bad and good clients. Other approach is creating the credit scoring model by based only on accepted clients who were eligible to get loan by decision technique of lender, however this method does not include rejected potential clients. In our study, our data allows us to use only second approach. There is no information about rejected people in data and therefore we will consider only accepted clients in our model. Our training data in this study is provided by one local company in Estonia.

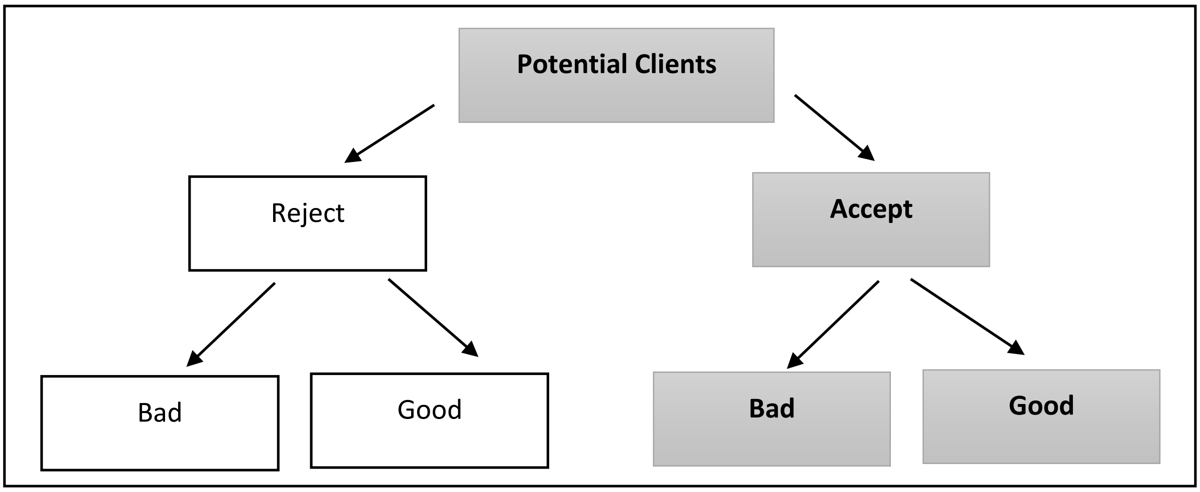

Figure 1: Components of Credit Scoring Model Population

Good/Bad Clients: Basically good clients are people who paid back their debts as scheduled and bad clients who did not pay back their debt as scheduled. Needless to say, bankruptcy and fraud is considered as bad clients.

Delinquency; occurs when a borrower fails to make a scheduled payment on a loan. Since loan payments are typically due monthly, lending industry customarily categorizes delinquent loans as either 30, 60, 90 or 120 or more days late depending on the length of time oldest unpaid loan payment has been overdue. [2]

Actual good client who paid their debts as scheduled however also some credit scoring systems may accept some problematic payment process as good client too. There are some situation where credit offices are able to consider them as good client even though they had problems to pay their debt as they scheduled:

- Borrower chooses to give the lender title of the property
- Lender agrees to renegotiate or modify the term of loan and forgives some or all of the delinquent principal and interest payments. (Loan modifications may take many form including change in the interest rate on the loan and extension of the length of the loan.) [2]

3.2 Variables

Variables	Explanation
Status *	Good clients(Y=1) who paid back their debts without facing any problem and bad clients(Y=0)
Sex	Female/Male
Age	Age of customers
Region	Countries such
Language	Mother tongue
Sum	Amount money is taken from bank (loan sum)
Period	Loan period in days
Income	Monthly income in EUR
Outcome	Monthly outcome in EUR
Family	Marriage Status
Education	Education level
WorkExperience	Clients who has more than one year experience and less than one year
Children	Number of Children
Estate	Number of real estate units
AlertsTotal	Total number of payment problems
AlertsActive	Number of active payment problems
AlertsClosed	Number of closed payment problems

*Dependent Variable

Table 1. Dependent and Independent Variables

Alerts Closed is highly correlated with Alerts Active (0.96) and therefore we did not include it in our model.

3.3 Classification of Regions

Estonia is divided into fifteen countries/regions. Capital city of Estonia is Tallinn and it is located in Harjumaa region. Second biggest city is Tartu and Tartu is located in Tartumaa region. Other countries are Pärnumaa, Järvamaa, Hiiuma, Ida-Virumaa, Valgamaa, Läänemaa, Saaremaa, Lääne-Virymaa, Viljandimaa, Jõgevamaa, Raplamaa, Põlvamaa and Võrumaa.

Region	Status
Harjumaa	,74
Tartumaa	,71
Pärnumaa	,71
Järvamaa	,70
Hiiumaa	,69
Ida-Virumaa	,69
Valgamaa	,69
Läänemaa	,68
Saaremaa	,68
Lääne-Virumaa	,65
Viljandimaa	,63
Jõgevamaa	,62
Raplamaa	,61
Põlvamaa	,60
Võrumaa	,58
Total	,71

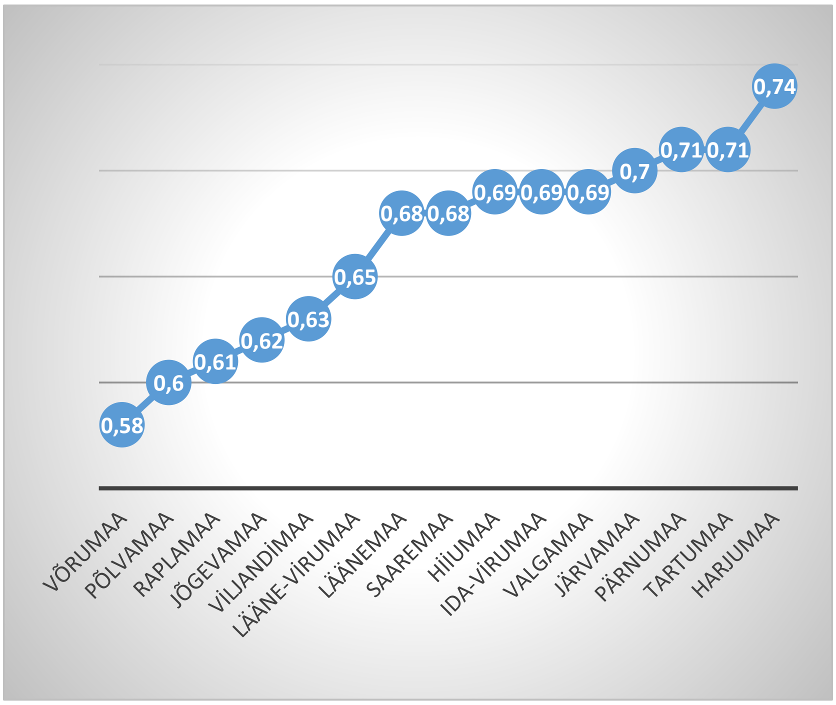

Figure 2. Means of Regions

Regions are reorganized by dendrogram of hierarchical clustering method of SPSS which helps to merge similar clusters. Clustering starts with every region in individual cluster and ends up with every region in one cluster regarding mean. We used same approach for age subgroups. This approach helps us to determine which regions should be merged at each step.

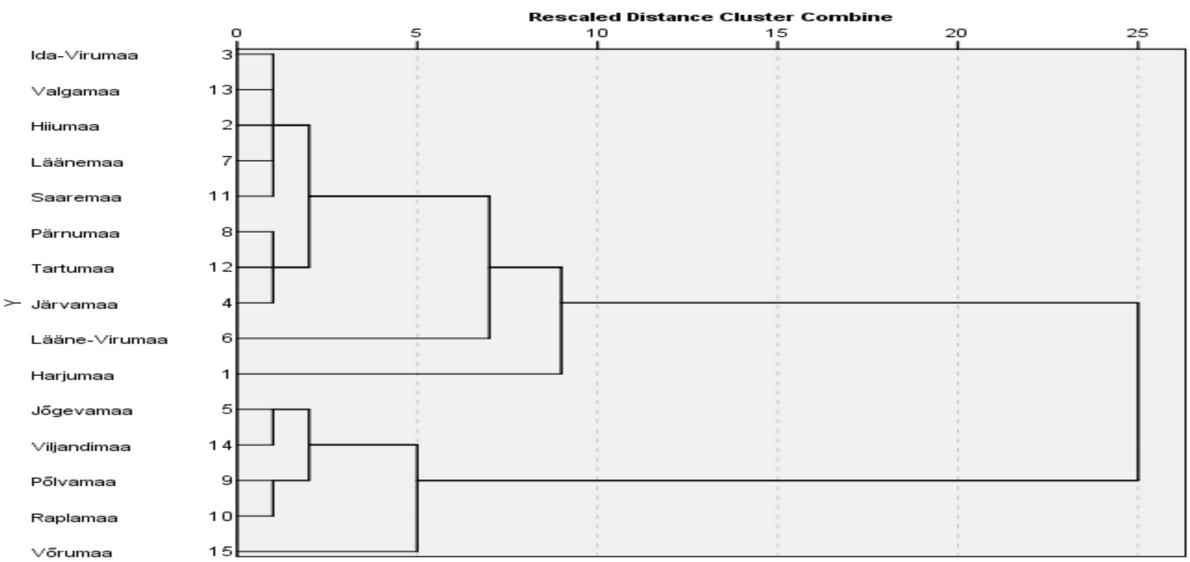

Figure 3. Countries Dendrogram

When distance indicator is 4, we have 5 regions groups by using clustering method. First region group is consisted of Harjumaa country. Region_1 group is consist of Tartumaa, Hiiuma, Ida-Virumaa, Valgamaa, Saarema, Läänem, Pärnumaa and Järvamaa countries. Region_2 group is consist of Lääne-Virumaa. Region_3 group is consisted of Põlvamaa, Raplamaa, Viljandimaa and Jõgevamaa countries and last region group is consisted of only Võrumaa. Probability of being good customer in these groups is 0.75, 0.70, 0.65, 0.62 and 0.58, respectively.

Regions	Status
Region_0	,74
Region_1	,70
Region_2	,65
Region_3	,62
Region_4	,58
Total	,71

Table 2. Means of Regions After Classification

3.4 Missing Values and Detection of Outliers

In Statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. [13]

There is four main ways to deal with missing values;

1. Exclude all data with missing values.
2. Exclude characteristics or records that have significant (more than 50%) missing values from the model, especially if the level of missing is expected to continue in the future.
3. The “missing values” can be treated as a separate attribute.
4. Impute missing values using statistical techniques. [6]

In statistics, an outlier is an observation point that is distant from other observations. [11] Outliers are values that fall outside of the normal range of value for a certain characteristic. These numbers may negatively affect the regression result, and are usually excluded. [6]

In our study, logistic regression requires complete data without missing cases and therefore we preferred first approach and we assumed that missing data is excluded. Also we preferred to use Box-plot method for detecting outliers. Box-plot method provides visual representation of dispersion of the data. This graphic has the lower quartile, Q_1, and upper quartile , Q_3 , with median (50th percentile). Upper and lower bounds are set a fixed distance with range between $Q_3 - Q_1$ .Upper and Lower fences to set at 1.5 times the interquartile range. Observations where are located out of these fences are potentially outliers/extreme values. One of the most important benefit of box-plot method is data does not have to be normally distributed.

4 Application For Modelling Credit Scoring By Logistic Regression

4.1 All Variables in the Model

At the beginning, adding all variables into model shows which variables are significant in credit scoring model. It gives opportunity to eliminate insignificant variables from our model in order to reach final model. It is important to observe relevant variables related to final model.

Original Value	Internal Value
Bad Client	0
Good Client	1

Table 3.Dependent Variables Encoding

In our model, status variable is dependent variable. Also status variable is binary variable, which is suitable for logistic regression, and it consists of only 0 and 1 inputs. Meaning of "0" is being "Bad Client" who is not likely to pay his/her debt and meaning of "1" is "being Good Client" who is likely to pay his/her debt properly as scheduled. We have 15 variables as our independent variables in our first model such as; sex, age, estate and marriage status… Sex, language, marriage status, work experience and regions are categorical variables and these variables expressed as dummy variables in our model.

		Frequency	Parameter coding			
			(1)	(2)	(3)	(4)
regions	,00	2912	,000	,000	,000	,000
	1,00	1960	1,000	,000	,000	,000
	2,00	214	,000	1,000	,000	,000
	3,00	365	,000	,000	1,000	,000
	4,00	65	,000	,000	,000	1,000
Language	Russian	2457	,000			
	Estonian	3059	1,000			
MarriageStatus	,00	2437	,000			
	1,00	3079	1,000			
workexpdum	,00	1425	,000			
	1,00	4091	1,000			
Sex	Female	2771	,000			
	Male	2745	1,000			

Table 4. Categorical Variables Coding

As we see on table, categorical variables consist one or more dummy variables. For example Sex has one dummy variable and being “Male” is represented by “Sex(1)” by IBM SPSS 21. Regions has more than one dummy variables and for example, region number 3 is represented by “Regions(3)”.

Let us see what result shows up when we use all of our variables in our first model.

Variables (a)	B (b)	S.E. ©	Wald (d)	df €	Sig. (f)	Exp(B) (h)
Sex(1)	-,325	,069	22,161	1	,000	,722
Language(1)	-,006	,066	,007	1	,934	,994
MarriageStatus(1)	,008	,069	,014	1	,907	1,008
workexpdum(1)	,206	,071	8,321	1	,004	1,229
regions			20,782	4	,000	
regions(1)	-,174	,069	6,326	1	,012	,840
regions(2)	-,403	,159	6,402	1	,011	,669
regions(3)	-,461	,128	13,012	1	,000	,631
regions(4)	-,550	,272	4,083	1	,043	,577
Age	,275	,040	48,071	1	,000	1,316
Sum	-,173	,048	12,902	1	,000	,841
Outcome	,184	,059	9,667	1	,002	1,202
Income	,019	,048	,153	1	,696	1,019
Period	,000	,000	,117	1	,733	1,000
Education	,205	,036	31,861	1	,000	1,227
Children	-,054	,038	2,020	1	,155	,947
Estate	,504	,051	95,736	1	,000	1,655
AlertsTotal	-,050	,014	11,969	1	,001	,951
AlertsActive	-,158	,042	13,932	1	,000	,854
Constant	,344	,115	8,956	1	,003	1,411

Table 5. All variables in the model

Briefly, review of table is:

- column (a) shows independents variables into model (In first model we add all variables into model),
- column (b) is Beta coefficients which give main component of regression model,
- (c) is Standard Errors are associated the beta coefficients and standard errors is required to use for estimating confidence intervals,

- column (d) provides Wald chi-square test values and column (f) provides two tailed p-values for each coefficients. Null hypothesis is $\beta_i = 0$ and alternative hypothesis is $\beta_i \neq 0$. If p-values is more than 0.05, 0.10, it means we cannot reject null hypothesis. It shows concerned variable is insignificant in model,
- (e) is the degrees of freedom for each of test coefficients,
- column (h) is odds ratios of predictors and these numbers are the exponentiation of the coefficients .

Interpretation of table: There are significant and insignificant variables in our first model. Categorical variable sex, work experience, regions are significant however language variable has 0.934 p-value and it means this variable is not significant when we regard 90%, 95% and 99% significant levels. Age, sum, outcome, education, estate, payment alert total and payment alert active are significant variables. On the other hand, incomes, period, number of children variables have 0.696, 0.733 and 0.155 p-values respectively and none of them are significant when we consider 90%, 95% and 99% significant level. Also it is possible to have further assessment on model by using backward method of SPSS 21. Backward method provides backward elimination which method starts with all variables in the model and eliminates variables step by step that are considered the least significant until all the remaining variable have a p-values below desired significant level. (you can check Annex 1)

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	6137,819	,086	,122

Table 6. R Square Estimates

There are two R^2 estimates in model summary stable. These are pseudo- R^2; meaning is these are analogous to R^2 in standard multiple regression, but do not carry the same interpretation. The Nagelkerke estimate is calculated in such a way as to be constrained between 0 and 1. So, it can be evaluated as indicating model fit; with a better model displaying a value closer 1. The larger Cox & Snell estimate is better the model; but it can be greater than 1.These metrics should be interpreted with caution, they offer little confidence in interpreting the model fit. [10]

In our case, our Nagelkerke R square is 0.122 and it is not so satisfied number and it is possible to increase the value by adding additional variables into model. For example, we can consider previous researches about credit scoring system related with regression analysis techniques. We can determine which additional parameters should be added to our model. Also – 2 log likelihood is 6137,819, if this value is smaller, it means model would be better.

	Observed		Predicted		
			Status		Percentage Correct
			Bad Client	Good Client	
Status	Bad Client		216	1377	13,6
	Good Client		174	3749	95,6
Overall Percentage					71,9

a. The cut value is ,500

Table 7. Classification Table

An intuitively appealing way to summarize the result of fitted logistic regression model is via a classification table. This table is the result of cross-classifying the outcome variable, y, with a dichotomous variable whose values are derived from the estimated logistic probabilities. In this application the coefficients produced by the model are used for predicting the outcome (in a binary way) rather that for estimating the probability of the event. To obtain the derived dichotomous variable we must define a cut-point and compare each estimated probability to cutoff point. If the estimated probability exceed cutoff point then we let the derived variable be equal to 1; otherwise it is equal to 0. The most commonly used value for cutoff is 0.5. [3]

Needless to say that all the modelling approaches may have errors and it happens in credit risk modeling too. There are two types of error in classification table that we can relate easily with credit scoring:

Error type 1 which means bad credits considered as good one,

Error type 2 which means good credits considered as bad one.

Errors type 1 is likely to cause losing the loan for lender. It may cause to increase risk of bankrupt however error type 2 is likely to cause losing the potentially good customers for bank. So error type 2 is missing opportunities. Our expectation is minimizing these two error types. In our first model, 1377 bad customers predicted as good customer which is error type I and 174 good customers are predicted as bad client which is error type 2.

4.2 Final Model After Elimination of Insignificant Variables

Firstly, we added all the variables into model in order to understand which variables are significant and insignificant in our model. To have final model, now we should apply the logistic regression by adding only significant variables. As we observed from 6th step of SPSS backward method output(Annex 1) that we have sex, work experience, regions, age, sum, outcome, education, estate, payment alert total and payment alert active as our significant independent variables.

	B	S.E.	Wald	df	Sig.	Exp(B)
Sex(1)	-,298	,065	21,368	1	,000	,742
workexpdum(1)	,209	,071	8,755	1	,003	1,232
regions			24,997	4	,000	
regions(1)	-,185	,068	7,369	1	,007	,831
regions(2)	-,415	,157	6,990	1	,008	,660
regions(3)	-,479	,123	15,226	1	,000	,619
regions(4)	-,572	,269	4,514	1	,034	,564
Age	,280	,039	51,464	1	,000	1,323
Sum	-,163	,040	16,383	1	,000	,849
Outcome	,182	,052	12,207	1	,000	1,200
Education	,212	,036	35,602	1	,000	1,236
Estate	,498	,051	94,986	1	,000	1,645
AlertsTotal	-,051	,014	12,824	1	,000	,950
AlertsActive	-,158	,042	13,915	1	,000	,854
Constant	,330	,093	12,465	1	,000	1,390

Table 8. Significant variables for model for all clients

As we see, coefficients are slightly changed when we include only significant variables into model. In our new model, we can see that sex dummy variable's coefficient number is -0.298 and it means "male" customer has negative effect on probability of being good customer. If customer have more than 1 year work experience, it contributes positively to being good customer. Also it is easy to see the effects of living in different regions on model. Based category region0 is represent only Harjumaa and customer from Harjumaa has better chance to be good candidate when we compare with other regions. Regions(1) contains; Tartumaa, Pärnumaa, Läänema and Valgamaa have slightly lower chance than Harjumaa region however regions(2) (involves; Ida-Virumaa, Viljandimaa, Järvamaa, and Hiiuma), regions(3) (involves; Raplaama, Lääne-Virumaa and Põlvamaa) and regions(4) (Saaremaa, Võrumaa and Jõgevamaa) have -0,415,-0,479 and -0,572 coefficients values respectively. Moreover, increasing age of customers, outcome, education level and number of estates contributes to increase probability of being good customer. On the other hand, sum, total and active alerts have negative impact. Needless to say that, when borrowed amount of money increases, credit-worthiness decreases at the same time.

And final model for probability of being good customer is:

$$ln\left(\frac{\pi}{1-\pi}\right) = 0{,}330 - 0{,}298 \cdot sex(1) + 0{,}209 \cdot workexp(1) - 0{,}195 \cdot region(1) - 0{,}415 \cdot region(2) - 0{,}479 \cdot region(3) - 0{,}572 \cdot region(4) + 0{,}29 \cdot age - 0{,}163 \cdot sum + 0{,}192 \cdot outcome + 0{,}212 \cdot education + 0{,}499 \cdot estate - 0{,}051 \cdot totpayalerts - 0{,}159 \cdot actpayalersts$$

As we can see, Cox & Snell R square has changed slightly (approximately 0,001) and Nagelkerke R square has exactly the same number. Removing insignificant variables from model did not change R square values.

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	6140,106[a]	,085	,122

Table 9. R squares for final model

Goodness of fit statistics assess the fit of a logistic model against actual outcomes. The goodness of fit test is the Hosmer Lemeshow (H-L) test. This statistics test H_0 hypothesis of HL test is

$$H_0: E[Y] = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}$$

using chi-square, and if becomes more than 0.05, shows that the model fits well to data.[7] Our p-value is 0.107 and p-value is bigger than 0.05 and in this case, we can say that model fits well enough to data.

Step	Chi-square	df	Sig.
1	13,144	8	,107

Table 10. Hosmer and Lemeshow Test

The Contingency Table for Hosmer and Lemeshow Test simply shows the observed and expected values for each category of the outcome variable as used to calculate the Hosmer and Lemeshow chi-square.

	Status = Bad Client		Status = Good Client		Total
	Observed	Expected	Observed	Expected	
1	291	299,097	261	252,903	552
2	231	239,700	321	312,300	552
3	229	209,633	324	343,367	553
4	181	185,120	371	366,880	552
5	170	162,675	380	387,325	550

6	144	141,841	407	409,159	551
7	125	121,589	427	430,411	552
8	117	102,461	435	449,539	552
9	64	80,969	487	470,031	551
10	41	49,915	510	501,085	551

Table 11. Contingency table for Hosmer and Lemeshow Test

A model classification table which describes both expected model classifications and actual model classifications. The Hosmer-Lemeshow table divides the data into 10 groups each representing the expected and observed frequency of both 1(Good Client) and 0(Bad Client) values. The expected frequency of data assigned to each deciles should match the actual frequency outcome and each deciles should contain data. [12]

Observed		Predicted		
		Status		Percentage Correct
		Bad Client	Good Client	
Status	Bad Client	216	1377	13,6
	Good Client	173	3750	95,6
Overall Percentage				71,9

a. The cut value is ,500

Table 12. Classification table

Classification table has almost same figures as classification table of our first model. Removing insignificant variables from model did not change the percentage of correct predictions. 71.9 percent of correct predictions may be acceptable as satisfying result however 1377 bad clients interpreted as good client which is increasing the risk of bankrupt (error type1). And we foresighted approximately only 14 percent of bad clients correctly (specificity) but needless to say that rejected borrowers by bank are not part of training dataset and proportion of clients are not equal (number of good clients is almost 3 times bigger). 95.6 percentage of correct predictions for good clients (sensitivity) is sufficiently high but 173 potential good customers are predicted as bad clients which shows that missing opportunities (error type 2).

Another point that can confirm the result of logit modelling is relative operating characteristic (ROC) curve, which shows a receiver operating characteristics and used for evaluating the logistic model as well. The ROC plot is merely the graph of points defined by sensitivity and 1-specificity. Customarily, sensitivity takes the y-axis and 1-specificity takes the x-axis. If the area under the curve becomes maximum amount, then the model fits data well. [7] The area under the ROC curve, which ranges from 0.5 to 1.0, provides a measure of the model's ability to discriminate between outcomes. [3]

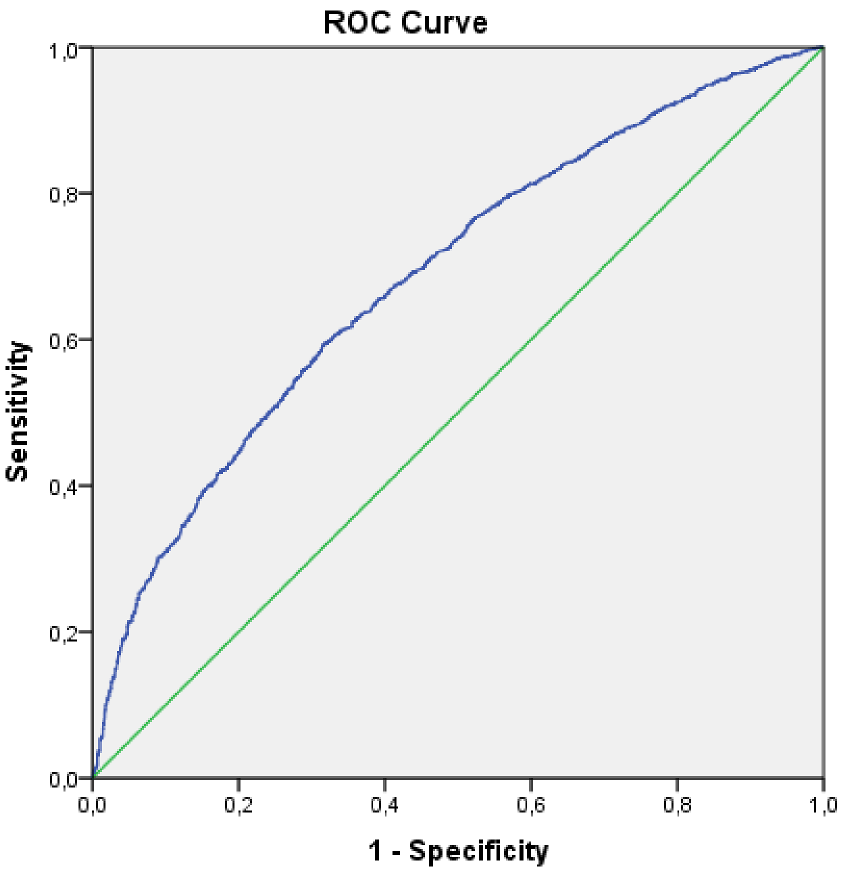

Figure 4. ROC Curve

SPSS outputs show that area under the curve is 0,685 with 95 percent of confidence interval (lower bound is 0,670 and upper bound is 0,700). Also the area under the curve has 0.00 p-value. It is significant and it means that logistic regression classifies the group significantly better than by chance.

Test Result Variable(s): Predicted probability

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,685	,008	,000	,670	,700

Table 13. Area under the curve

Decision of cutting point is also important and it depends on policies of lenders. They would prefer to take increasing risk of bankrupt while they add new customers or they would prefer to decrease the risk of bankrupt while they lose potentially good clients. In this project, our goal was having maximum percentage of correct prediction from classification table. And as we can observe from table below that best cut-off point is 0.50 in order to have higher percentage of correct predictions.

Classification Table					
Cut-off points	Bad Client	Error Type I	Error Type II	Good Client	Percentage Correct
0,3	10	1583	2	3921	71,3%
0,4	50	1543	34	3889	71,4%
0,45	114	1479	78	3845	71,8%
0,5	**216**	**1377**	**173**	**3750**	**71,9%**
0,55	355	1238	338	3585	71,4%
0,6	537	1056	603	3320	69,9%
0,7	995	598	1422	2501	63,4%
0,8	1386	207	2539	1384	50,2%

Table 14. Classification table for different cut-off points

Additionally alternative way to decide cut-off point is considering specificity and sensitivity table (Annex 2). According to table, cut-off point is 0,701 with 0,633 sensitivity and 0,367 1-specifity. This approach provides balanced result which has 63 percentage of correct prediction for good and bad clients separately.

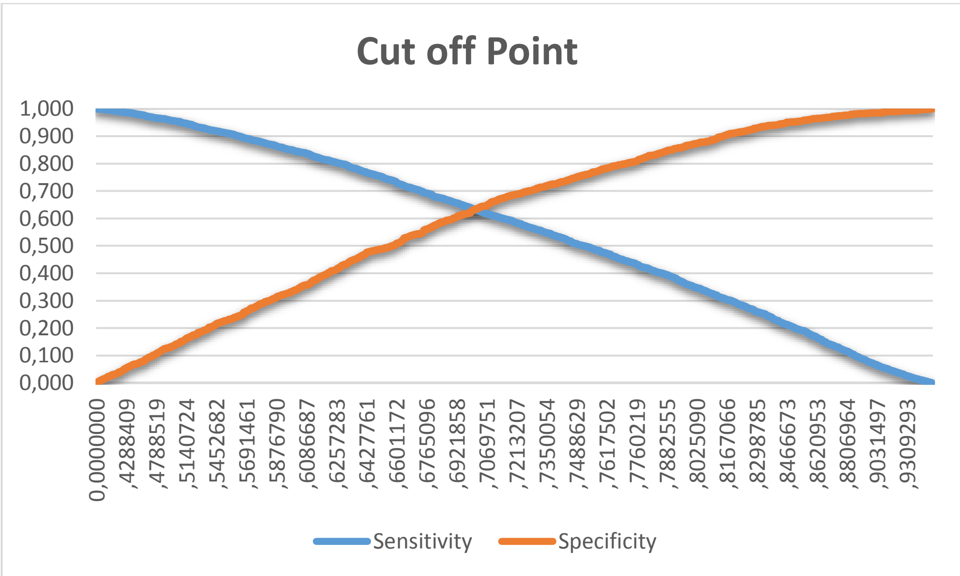

Figure 5. Sensitivity and specificity graph

5 Segmentations/Models for 4 Different Age Groups

Following overview is based on [6]

In some cases, using several scorecards for a portfolio provides better risk differentiation than using one scorecard on everyone. This is usually the case where a population is made up of distinct subpopulations, and where one scorecard will not work efficiently for all of them (i.e., we assume that different subpopulations in our portfolio). The process of identifying these subpopulations is called segmentation. There are two main ways in which segmentation can be done:

1. Generating segmentation ideas based on experiences and industry knowledge, and then validating these ideas using analytics
2. Generating unique segments using statistical techniques such as clustering or decision trees.

In either case, any segments selected should be large enough to enable meaningful sampling for separate scorecard development. Segments that exhibit distinct risk performance, but have insufficient volume for separate scorecard development, can still be treated differently using different cut-offs or other strategy considerations.

Segmentation, whether using experience or statistical methods, should also be done with future plans in mind. Most analysis and experience is based on the past, but scorecards need to be implemented in the future, on future applicant segments. One way to achieve this is by adjusting segmentation based on, for example, the organization's intended target market. Traditionally, segmentation has been done to identify an optimum set of segments that will maximize performance—the approach suggested here is to find a set of segments for which the organization requires optimal performance, such as target markets. This approach underscores the importance of trying to maximize performance where it is most needed from a business perspective and ensures that the scorecard development process maximizes business value. This is an area where marketing staff can add value and relevance to scorecard development projects.

Typical segmentation areas used in the industry include those based on:

• **Demographics.** Regional (province/state, internal definition, urban/rural, postal-code based, neighbourhood), age, lifestyle code, time at bureau, tenure at bank

• **Product Type.** Gold/platinum cards, length of mortgage, insurance type, secured/unsecured, new versus used leases for auto, size of loan

• **Sources of Business (channel).** Store-front, take one, branch, Internet, dealers, brokers

• **Data Available.** Thin/thick (thin file denotes no trades present) and clean/dirty file (dirty denotes some negative performance) at the bureau, revolver/transactor for revolving products, SMS/voice user

• **Applicant Type.** Existing/new customer, first time home buyer/mortgage renewal, professional trade groups (e.g., engineers, doctors,etc.)

• **Product Owned.** Mortgage holders applying for credit cards at the same bank.

In our study, we will create 4 different segmentation groups based one age independent variable. Age variable has 4 subgroups; clients who are "30 years old and younger than 30", clients "between 31 and 45", "clients between 45 and 60" and "clients who are older than 60 years old". These groups have 1553, 2244, 1453 and 341 observations respectively.

	Frequency	Percent
30 And Younger Than 30 Years Old	1553	27,8
Older Than 30 And Younger Than 46	2244	40,1
Older Than 45 And Younger Than 60	1453	26,0
Older Than 60 Years Old	341	6,1
Total	5591	100,0

Table 15. Age groups

5.1 Model for 30 And Younger Than 30 Years old Clients

We have 1553 clients who are 30 and younger than 30 years old however 1531 of clients in analysis because 22 of them are missing cases.

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	1531	98,6
	Missing Cases	22	1,4
	Total	1553	100,0
Unselected Cases		0	,0
Total		1553	100,0

Table 16. Case summary

Result shows quite different model if we compare with our final model without segmentation. Regions, outcome, income, education and total alerts are significant. Language is also significant with 90% of significance level. Output shows us that clients whose mother tongue is Estonian are likely to be more trustworthy borrowers. Needless to say, number of total alerts has negative effect on creditworthiness as we can foresee. Number of estate and higher education are contributing to be good customer.

	B	S.E.	Wald	df	Sig.	Exp(B)
Language(1)	,199	,117	2,918	1	,088	1,221
regions			33,162	4	,000	
regions(1)	-,488	,132	13,690	1	,000	,614
regions(2)	-,746	,212	12,336	1	,000	,474
regions(3)	-,936	,343	7,455	1	,006	,392
regions(4)	-1,339	,498	7,217	1	,007	,262
Outcome	,297	,113	6,921	1	,009	1,345
Income	-,161	,077	4,413	1	,036	,851
Education	,257	,060	18,191	1	,000	1,293
Estate	,408	,120	11,501	1	,001	1,504
AlertsTotal	-,087	,026	10,966	1	,001	,917
Constant	,362	,159	5,196	1	,023	1,437

Table 17. Variables into model for first age group

And in this age group, model for probability of being good customer is:

$$\ln\left(\frac{\pi}{1-\pi}\right) = 0{,}362 + 0{,}199 \cdot language(1) - 0{,}488 \cdot region(1) - 0.746 \cdot region(2) - 0{,}936 \cdot region(3) - 1{,}339 \cdot region(4) + 0{,}297 \cdot outcome - 0{,}161 \cdot income + 0{,}257 \cdot education + 0{,}408 \cdot estate - 0{,}087 \cdot totpayalerts$$

Model provides 62.4 percentage of correct predictions when cut-off point is 0.5. Good clients are well predicted however correct prediction for bad clients are low and around 29 percentage.

	Observed		Predicted		
			Status		Percentage Correct
			Bad Client	Good Client	
	Status	Bad Client	183	440	29,4
		Good Client	136	772	85,0
	Overall Percentage				62,4

a. The cut value is ,500

Table 18. Classification table

There are some classification results when we determined some specific different cut-off points. It is possible to get better result when cut-off point is 0.55 than our cut-off point is 0.5.

Classification Table					
Cut-off points	Bad Client	Error Type I	Error Type II	Good Client	Percentage Correct
0,3	15	608	4	904	60%
0,4	64	559	28	880	61,7%
0,5	183	440	136	772	62,4%
0,55	**272**	**351**	**210**	**698**	**63,4%**
0,6	381	242	383	525	59,2%
0,7	549	74	714	194	48,5%

Table 19. Classification table for different cut-off points

SPSS outputs shows that area under the curve is 0,636 with 95 significance level (lower bound is 0,608 and upper bound is 0,665). Also the area under the curve has 0.00 p-value. It express that logistic regression classifies the group significantly better than by chance.

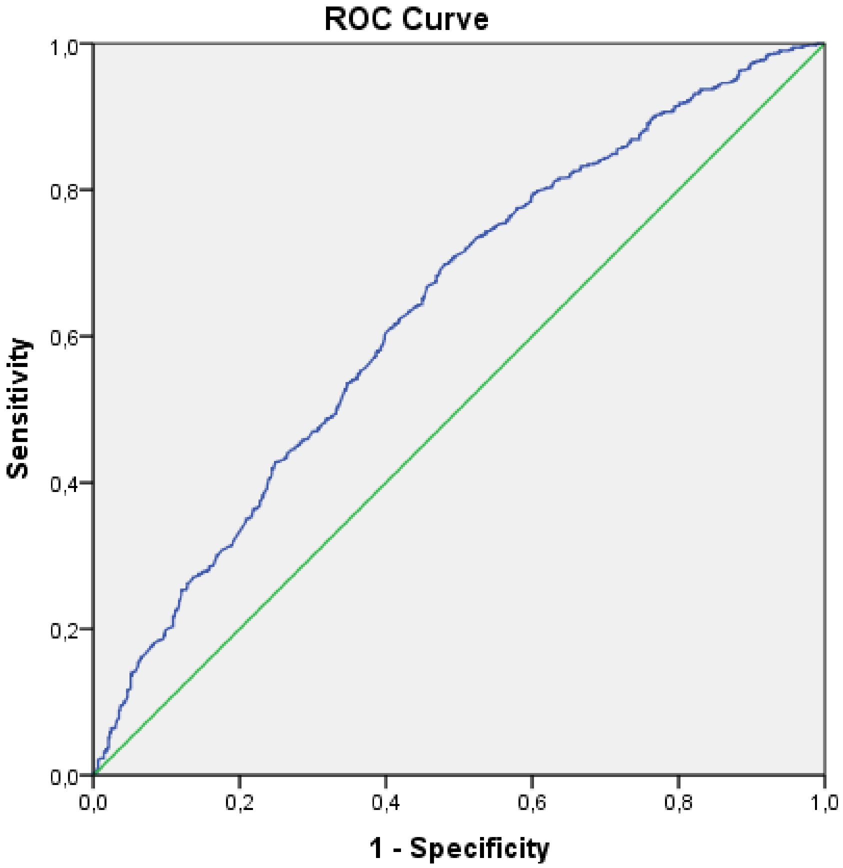

Figure 6. Roc Curve

Test Result Variable(s): Predicted probability

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,636	,014	,000	,608	,665

Table 20. Area under the curve

5.2 Model For Between 31 To 45 Years Old

Let us check the next age group which is consist of clients who are older 30 years old and younger than 46 years old. We have 2244 clients in this age group however 1.3 percent of clients is missing cases. 2215 of clients is included in analysis.

Unweighted Cases		N	Percent
Selected Cases	Included in Analysis	2215	98,7
	Missing Cases	29	1,3
	Total	2244	100,0
Unselected Cases		0	,0
Total		2244	100,0

Table 21. Case summary

We have 8 significant variables in our model. These variables are sex, work experience, region, outcome, period, education level, number of estates and active alerts. Having more than one year work experience is contributing creditworthiness positively. Moreover client who have higher education level and number of estates or ownership of estate(s) are increasing probability of being good customer. On the other hand, active alerts is reducing the probability of being good customer. Also period has negative effect too and we can say that longer instalment period reduces chance to be good client for candidate. Coefficient of period is so small because of scale problem. Additionally, female candidates are more reliable clients than man in this age group.

	B	S.E.	Wald	df	Sig.	Exp(B)
Sex(1)	-,343	,102	11,256	1	,001	,710
workexpdum(1)	,349	,116	9,143	1	,002	1,418
regions(1)	-,643	,271	5,619	1	,018	,526
Outcome	,177	,074	5,709	1	,017	1,194
Period	-,001	,000	4,203	1	,040	,999
Education	,228	,056	16,809	1	,000	1,256
Estate	,535	,076	50,086	1	,000	1,707
AlertsActive	-,236	,059	15,698	1	,000	,790
Constant	,284	,148	3,688	1	,055	1,329

Table 22. Variables in the equation for second age group

Clients who live in Põlvamaa and Viljandimaa are to reduce probability of being customer. Living in rest of countries provides candidates some benefits in this age group model. This age group's specific model is simpler and plainer than previous models and model for probability of being good customer is:

$$\ln\left(\frac{\pi}{1-\pi}\right) = 0{,}284 - 0{,}343 \cdot sex(1) - 0{,}643. region(1) + 0{,}177 \cdot outcome - 0{,}001 \cdot period + 0{,}228 \cdot education + 0{,}535 \cdot estate - 0{,}236 \cdot actpayalerts$$

Model in this age group provides 72.1 percentage of correct predictions when cut-off point is 0.5. Good clients are predicted with 97.7 percentage correctly however model is not successful to predict bad clients efficiently and percentage of correct predictions is very low.

Observed		Predicted		
		Status		Percentage Correct
		Bad Client	Good Client	
Status	Bad Client	43	583	6,9
	Good Client	36	1553	97,7
Overall Percentage				72,1

a. The cut value is ,500

Table 23. Classification table

Also there are some specific cut-off points and percentage of correct predictions for these cut-off points. Highest percentage of correct predictions in classification table is 0.5 in this specific age group. It is possible to increase percentage of correct prediction for bad clients' predictions however it has negative impact on general model which decreases percentage of correct prediction dramatically.

Classification Table					
Cut-off points	Bad Client	Error Type I	Error Type II	Good Client	Percentage Correct
0,4	14	612	12	1577	71,8%
0,45	25	601	23	1566	71,8%
0,5	**43**	**583**	**36**	**1553**	**72,1%**
0,55	84	542	84	1505	71,7%
0,6	146	480	180	1409	70,2%
0,7	375	251	582	1007	62,4%

Table 24. Classification table for different cut-off points

Area under the curve is 0,664 with 95 significance level and confidence interval lower bound is 0,639 and upper bound is 0,688. If area under the curve is between 0.6 and 07, we can say it is on acceptable level. Also the area under the curve has 0.00 p-value. It expresses that logistic regression classifies the group significantly better than by chance.

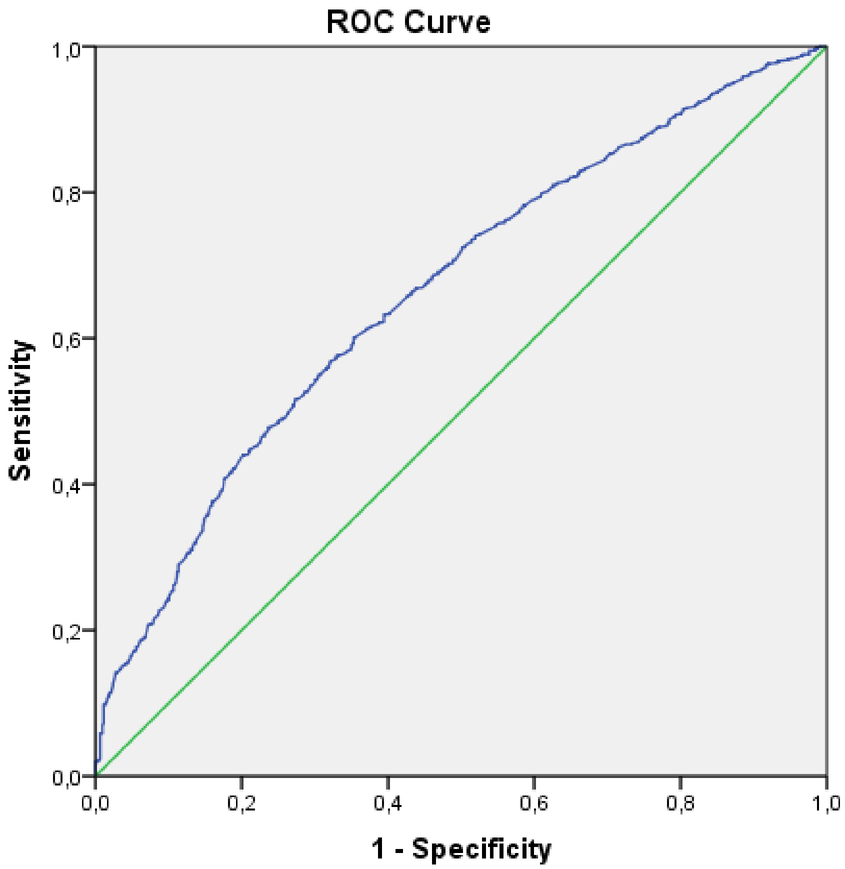

Figure 7. ROC Curve

Test Result Variable(s): Predicted probability

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,664	,012	,000	,639	,688

Table 25. Area under the curve

5.3 Model For Between 46 To 60 Years Old

In this study, we have 1453 clients who are between 46 and 60 years old. 19 of them are missing value and they are excluded. 1434 clients are included in analysis.

Unweighted Cases		N	Percent
Selected Cases	Included in Analysis	1434	98,7
	Missing Cases	19	1,3
	Total	1453	100,0
Unselected Cases		0	,0
Total		1453	100,0

Table 26. Case summary

Let us check the output. Model have 8 significant variables. Language dummy variable is part of model when we consider 90% significance level but it is possible to exclude it when we take significance level 95%. Significant variables are sex, language, region, sum, income, education, estate and active alerts. Result shows us that female customers are more reliable than men customers in this age group like previous one. Also it is easy to see that living in different countries have impact on being good customer. Amount of borrowed money is also important. It trigger bigger risk to pay it back while amount of sum is increasing. Higher education level and number of estates are contributing to be good customer. Customer, who have higher education level and have his/her own estate, is likely to be more trustworthy candidate.

	B	S.E.	Wald	df	Sig.	Exp(B)
Sex(1)	-,376	,141	7,080	1	,008	,686
Language(1)	-,264	,143	3,382	1	,066	,768
regions			14,494	2	,001	
regions(1)	-,577	,210	7,542	1	,006	,562
regions(2)	-1,317	,463	8,092	1	,004	,268
Sum	-,264	,089	8,832	1	,003	,768
Income	,219	,093	5,568	1	,018	1,244
Education	,175	,083	4,447	1	,035	1,191
Estate	,447	,088	25,982	1	,000	1,563
AlertsActive	-,265	,078	11,403	1	,001	,767
Constant	1,032	,207	24,769	1	,000	2,808

Table 27. Variables in the model

And in this age group, the model looks like this:

$$ln\left(\frac{\pi}{1-\pi}\right) = 1{,}032 - 0{,}376 \cdot sex(1) - 0.264 \cdot language - 0{,}577 \cdot region(1) - 1{,}317 \cdot region(2) - 0{,}264 \cdot sum + 0{,}219 \cdot education + 0{,}447 \cdot estate - 0{,}265 \cdot actpayalerts$$

Logistic regression model provides 79.6 percent of correct prediction when cut-off point is 0.5. Good clients are well predicted with 99 percentage however bad clients is predicted with only 5 percentage.

	Observed		Predicted		
			Status		Percentage Correct
			Bad Client	Good Client	
	Status	Bad Client	14	282	4,7
		Good Client	11	1127	99,0
	Overall Percentage				79,6

a. The cut value is ,500

Table 28. Classification Table

As we can see on classification table below, current cut-off point provides higher percentage of correct predictions than other cut-off points. It is possible to increase percentage of correct predictions to determine bad clients if we pick higher cut-off point however unfortunately it causes another problem which will effect on general percentage of correct prediction sharply.

Classification Table					
Cut-off points	Bad Client	Error Type I	Error Type II	Good Client	Percentage Correct
0,4	7	289	4	1134	79,5
0,45	8	288	6	1132	79,5
0,5	**14**	**282**	**11**	**1127**	**79,6**
0,55	15	281	17	1121	79,2
0,6	27	269	38	1100	78,6
0,7	85	211	116	1022	77,2

Table 29. Classification table for different cut-off points

Area under the curve is 0,681 with 95 significance level and confidence interval for lower bound is 0,647 and upper bound is 0,715. Area under the curve is around 0.7. We can say that model has ability to discriminate between two outcomes. It is possible to interpret the area under the curve is good enough when it is around 0.7. Also the area under the curve has 0.00 p-value. It means that logistic regression classifies the group significantly better than by chance.

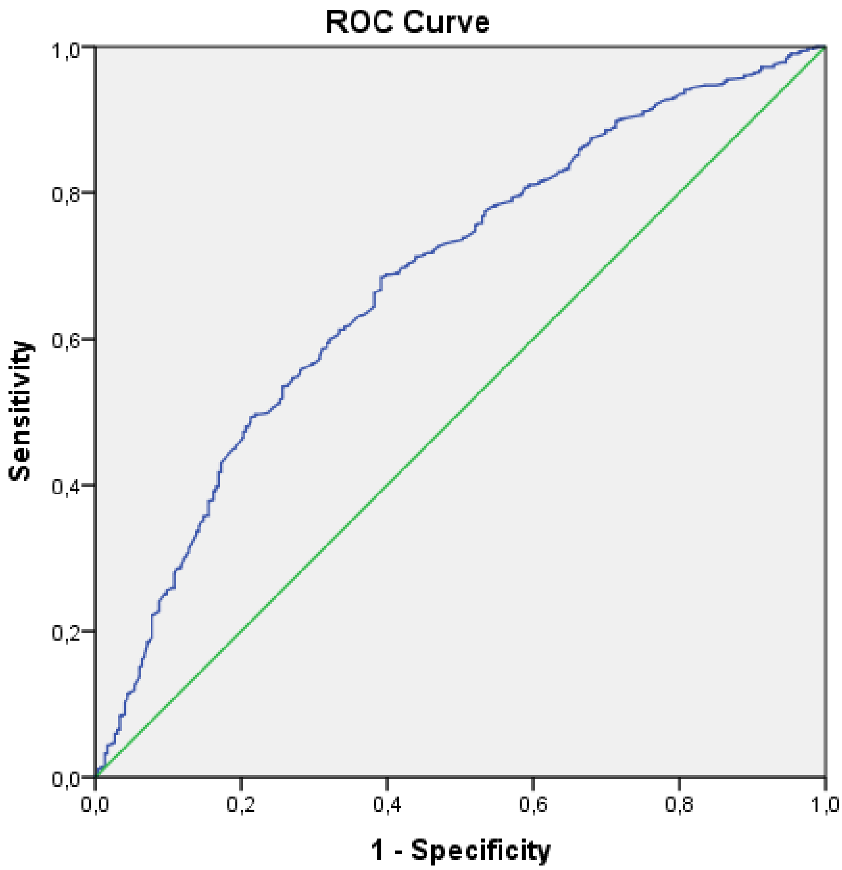

Figure 8. ROC curve

Test Result Variable(s): Predicted probability

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,681	,017	,000	,647	,715

Table 30. Area under the curve

5.4 Model For Older Than 60 Years Old Clients

There are 341 clients who are older than 60 years old. Model includes 336 of them because of 5 of them have missing values. Logistic regression requires complete data without missing cases. Therefore we excluded missing cases.

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	336	98,5
	Missing Cases	5	1,5
	Total	341	100,0
Unselected Cases		0	,0
Total		341	100,0

Table 31. Case summary

We have the simplest and the plainest model in this age group and the latest model has only 4 significant variables. These variables are region, sum, number of estates and total alerts respectively. It is easy to see that if client demand to borrow bigger amount of money, it will increase the risk of paying debt back and it will decrease creditworthiness of candidate in our model for this age group. Needless to say that total alerts have negative impact as we can presume.

	B	S.E.	Wald	df	Sig.	Exp(B)
regions(1)	-2,698	1,048	6,627	1	,010	,067
Sum	-,671	,224	8,933	1	,003	,511
Estate	1,228	,311	15,605	1	,000	3,413
AlertsTotal	-,317	,100	10,116	1	,001	,728
Constant	4,371	1,063	16,890	1	,000	79,100

Table 32. Variables in the model

Also we can see that region variable is important indicator. Living in Jõgevamaa, Läänemaa, Põlvamaa, Valgamaa, Võrumaa and Tartumaa are contributing more to be good customer than other countries in this age group which is consisted of clients who are older than 60 years old.

And the model is as follows:

$$\ln\left(\frac{\pi}{1-\pi}\right) = 4{,}371 - 2{,}698 \cdot region(1) - 0{,}671 \cdot sums + 1{,}228 \cdot estate - 0{,}317 \cdot totpayalerts$$

Classification table shows us that model provides 86.2 percent of correct predictions when cut-off point is 0.5. 291 of 293 good clients are predicted correctly (99.3%) however bad clients are predicted with only 6.3 percentage. Therefore model is able to predict 3 of 48 bad clients correctly.

	Observed		Predicted		
			Status		Percentage Correct
			Bad Client	Good Client	
	Status	Bad Client	3	45	6,3
		Good Client	2	291	99,3
	Overall Percentage				86,2

a. The cut value is ,500

Table 33. Classification table

According to classification table below, we can say that model provides 87.1 percentage of correct predictions when cut-off point is 0.6 which gave better result than when cut-off point is 0.5.

Classification Table					
Cut-off points	Bad Client	Error Type I	Error Type II	Good Client	Percentage Correct
0,4	0	48	1	292	85,6%
0,45	2	46	2	291	85,9%
0,5	3	45	2	291	86,2%
0,55	6	42	4	289	86,5%
0,6	**12**	**36**	**8**	**285**	**87,1%**
0,7	17	31	13	280	87,1%
0,75	28	20	44	249	81,2%

Table 34. Classification table for different cut-off points

Area under the curve is 0,775 and confidence interval with 95 significance level is from 0,699 to 0,851. Since area under the curve is between 70% and 80%, it is a good model. Moreover we can say that ability of a fitted model to discriminate between two outcomes is on fairly acceptable level. Also, the area under the curve has 0.00 p-value. It means that logistic regression classifies the group significantly better than by chance.

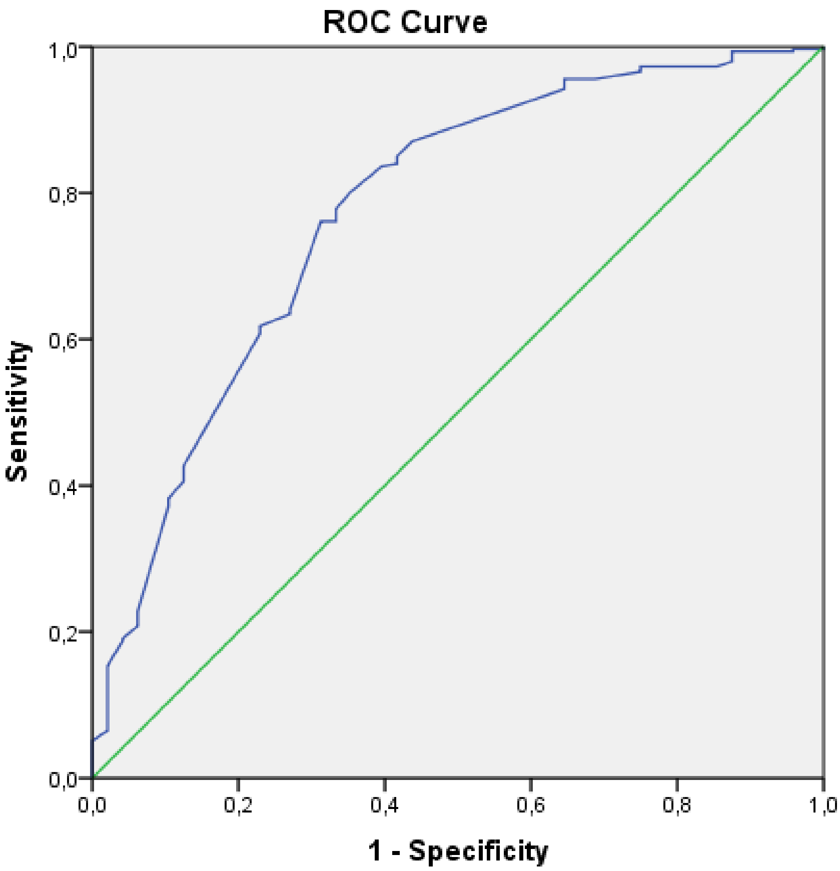

Figure 9. ROC curve

Area Under the Curve

Test Result Variable(s): Predicted probability

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,775	,039	,000	,699	,851

Table 35. Area under the curve

6 Results

Firstly, we created one model for all clients without any separation which the model is created by logistic regression and according to our model, sex, work experience, regions, age, sum, outcome, education, estate, total alerts and active alerts are significant variables and model provided 71.9% of correct prediction overall when cut-off point is 0.5.

In second part, we separated clients regarding their age groups in order to have different credit scoring model for different age group. Purpose of creating different models for each group is observing whether it helps increasing percentage of correct prediction or not which may be dramatically critical in real life for lenders. Therefore we divided data four different subgroups regarding each age groups and we created 4 different models for each age groups. These are clients who are 30 years old and younger, clients who are between 31 and 45 years old, clients who are between 46 and 60 years old and clients who are older than 60 years old respectively. In first model for age groups, significant variables are language, region, outcome, income, education, estate and total alerts and model provided 63.4% of correct predictions. Second model for age groups have 8 significant variables; sex, work experience, region, outcome, period, education, estate and active alerts. Second segmentation age group provided 72.1% of correct prediction. Next model have sex, language, regions, sum, income, education, estate and active alerts as significant independent variables which provides 79.6% of correct prediction. Last model for age groups has only 4 significant variables; region, sum, estate and total alerts and clients is predicted with 87.1% correctly by this model.

	Classification Table					
	Cut-off points	Bad Client	Error Type I	Error Type II	Good Client	Percentage
Final Model (one model)	**0,5**	**216**	**1377**	**173**	**3750**	**71,9%**
30 and Younger	0,55	272	351	210	698	63,4%
Between 31 to 45	0,5	43	583	36	1553	72,1%
Between 46 to 60	0,5	14	282	11	1127	79,6%
Older than 60	0,6	12	36	8	285	87,1%
4 Segmentation models together		341	1252	265	3663	**72,5%**

Table 36. Model comparison between model for all clients and segmentation

We predicted 72.5% of our clients correctly with helps of segmentation regarding age variable. And as we mentioned before, only one model for all clients provided 71.9% of correct prediction. Segmentation helped to increase correct predictions from 71.9% to 72.5% and it is equal to additionally 38 correctly predicted customers in our study. Error type I decreased from

1377 to 1252 and it implies that number of bad customers considered as good customer which cause to rise risk of bankrupt and frauds. On the other hand, it caused to increase error type II from 173 to 265 which implies that number of client who are actually good but we considered as bad clients.

Conclusion

In this master thesis, we showed that logistic regression method is one of the efficient approaches to build credit scoring models for small loans. Our first model predicts customers' behavior with 71.9 percentage correctly. We can say that one model for all clients provides fairly good predictions. However, one single model does not work efficiently for all different customers and it is possible to improve accuracy of predictions by creating different models for subgroups. Segmentation models regarding subgroups of age variable provides us 72.5 percentage of correct predictions and it means additionally 38 correctly predicted clients which may be considerably important for lenders. It may save thousands of euros worth loss of money. Also this additional information may help borrowers too. Hereby improved model can protect borrowers from liabilities of debt which borrowers are not able to handle and pay back in the future. Obviously even one more correct prediction makes difference in this sector and segmentation is great technique to improve credit scoring models. From our perspective, it is definitely worth spending time on it, in order to improve the credit scoring model. Moreover our models showed that having number of estates and high education increase the probability of being good customer among candidates. These two variables are significant characteristics of models. Also sex and language dummy variables are considerably important. Results suggest that female candidates are more reliable than male candidates and clients whose mother tongue is Estonian are more reliable as well. Work experience, income, outcome, alerts and age should be considered notably important variables. And one model for all clients showed that older clients are likely to be trustworthy customers.

All in all, our final comment is creating credit scoring model by using logistic regression can help to decrease the risk of losing money and it will help to find new customers who are more reliable. Obviously lenders should consider more than one technique to decide whether candidate is creditworthy or not. Most of professionals suggest that lenders must contribute to decision process with their knowledge, training and experience while credit scoring model proceeds. Combination of credit scoring model and experts' foresight helps to increase effectiveness. For example, if potential customer's probability of being good client is 0.47 and it is around cut-off point (assume that cut-off point is 0.5) and basically it means customer is in grey area. Client is considered as bad client by model but potentially client would be good customer indeed. Hereby expert may provide good consultation to discriminate good and bad customer in this grey area where probability is +/-0.05 around cut-off point. On the other hand, segmentation is very useful to improve model's effectiveness and in my opinion it is possible to improve credit scoring model more by segmentation if we create more models regarding subgroups which is provided by statistical methods like decision trees and clustering.

References

[1] Anderson, R., **The Credit Scoring Toolkit; Theory and Practice for Retail Credit Risk Management and Decision Automation**, 2007, Oxford University Press, p.3-5.

[2] Avery, R.B., Bostic, R.W., Calem, P.S., **Credit Risk, Credit Scoring and the performance of home mortgage**, Federal Reserve Bulletin, July 1996, p.2-8.

[3] Hosmer, D.W., Lemeshow, S., Sturdivant, R.X., **Applied Logistic Regression**, Third Edition, 2013, Wiley, p.1, 169-175.

[4] Menard, S,. **Applied Logistic Regression Analysis,** Second Edition, 2002, Sage University Paper, p.6.

[5] Peng, C.Y., Lee, K.L., Ingersoll, G.M., **An Introduction to Logistic Regression Analysis and Reporting**, 2002, Indiana University.

[6] Siddiqi, N., **Credit Risk Scorecards; Developing and Implementing Intelligent Credit Scoring**, 2006, John Wiley & Sons inc, p.1, 46-49, 76-92.

[7] Soureshjani, M.H., Kimiagari, A.M., **Calculating Best Cut-Off Point Using Logistic Regression And Neural Network On Credit Scoring Problem**, AJ of Business management, 2013,Amirkabir University of Technology, p.1416-1419.

[8] Thomas, L.C., Edelman, D.B., Crook, J.N.,**Credit Scoring and Its Applications**, 2002, Society for Industrial and Applied Mathematics, p.2-10, 41.

[9] Wright State University; Lecture Notes, <http://www.wright.edu/~thaddeus.tarpey/ES714glm.pdf> [access: 3.02.2015]

[10] University of North Texas, Research and Statistical Support: module 9, http://www .unt.edu/rss/class/Jon/SPSS_SC/Module9/M9_LogReg/SPSS_M9_LogReg.htm, [access : 27.03.2016]

[11] Engineering Statistics Handbook, <http://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm > [access: 2.03.2016]

[12] <http://appricon.com/index.php/logistic-regression-analysis.html> [access: 21.03.2016]

[13] <http://www.encepp.eu/standards_and_guidances/methodologicalGuide5_3.shtml> [access: 17.03.2016]

Appendix

Annex1: One model for everyone stepwise backward method

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Sex(1)	-,325	,069	22,161	1	,000	,722
	Language(1)	-,006	,066	,007	1	,934	,994
	MarriageStatus(1)	,008	,069	,014	1	,907	1,008
	workexpdum(1)	,206	,071	8,321	1	,004	1,229
	regions			20,782	4	,000	
	regions(1)	-,174	,069	6,326	1	,012	,840
	regions(2)	-,403	,159	6,402	1	,011	,669
	regions(3)	-,461	,128	13,012	1	,000	,631
	regions(4)	-,550	,272	4,083	1	,043	,577
	Age	,275	,040	48,071	1	,000	1,316
	Sum	-,173	,048	12,902	1	,000	,841
	Outcome	,184	,059	9,667	1	,002	1,202
	Income	,019	,048	,153	1	,696	1,019
	Period	,000	,000	,117	1	,733	1,000
	Education	,205	,036	31,861	1	,000	1,227
	Children	-,054	,038	2,020	1	,155	,947
	Estate	,504	,051	95,736	1	,000	1,655
	AlertsTotal	-,050	,014	11,969	1	,001	,951
	AlertsActive	-,158	,042	13,932	1	,000	,854
	Constant	,344	,115	8,956	1	,003	1,411
Step 2[a]	Sex(1)	-,325	,069	22,154	1	,000	,722
	MarriageStatus(1)	,008	,069	,014	1	,905	1,008
	workexpdum(1)	,206	,071	8,331	1	,004	1,229
	regions			22,714	4	,000	
	regions(1)	-,174	,069	6,377	1	,012	,840
	regions(2)	-,405	,157	6,630	1	,010	,667
	regions(3)	-,464	,124	14,036	1	,000	,629
	regions(4)	-,553	,270	4,191	1	,041	,575
	Age	,275	,040	48,065	1	,000	1,316
	Sum	-,173	,048	12,894	1	,000	,841
	Outcome	,183	,059	9,660	1	,002	1,201
	Income	,018	,048	,148	1	,701	1,019
	Period	,000	,000	,116	1	,734	1,000
	Education	,205	,036	32,048	1	,000	1,228
	Children	-,054	,038	2,024	1	,155	,947

	Estate	,504	,051	95,710	1	,000	1,655
	AlertsTotal	-,050	,014	12,051	1	,001	,951
	AlertsActive	-,158	,042	13,933	1	,000	,854
	Constant	,342	,112	9,344	1	,002	1,408
Step 3[a]	Sex(1)	-,326	,069	22,445	1	,000	,722
	workexpdum(1)	,206	,071	8,347	1	,004	1,229
	regions			22,731	4	,000	
	regions(1)	-,174	,069	6,373	1	,012	,840
	regions(2)	-,405	,157	6,642	1	,010	,667
	regions(3)	-,464	,124	14,053	1	,000	,629
	regions(4)	-,553	,270	4,188	1	,041	,575
	Age	,275	,039	48,654	1	,000	1,317
	Sum	-,173	,048	12,880	1	,000	,841
	Outcome	,183	,059	9,648	1	,002	1,201
	Income	,018	,048	,149	1	,699	1,019
	Period	,000	,000	,116	1	,734	1,000
	Education	,205	,036	32,036	1	,000	1,228
	Children	-,053	,036	2,164	1	,141	,949
	Estate	,504	,051	96,027	1	,000	1,655
	AlertsTotal	-,050	,014	12,037	1	,001	,951
	AlertsActive	-,158	,042	13,942	1	,000	,854
	Constant	,345	,109	10,034	1	,002	1,412
Step 4[a]	Sex(1)	-,327	,069	22,672	1	,000	,721
	workexpdum(1)	,206	,071	8,335	1	,004	1,229
	regions			22,616	4	,000	
	regions(1)	-,173	,069	6,293	1	,012	,841
	regions(2)	-,404	,157	6,611	1	,010	,668
	regions(3)	-,461	,123	13,951	1	,000	,630
	regions(4)	-,551	,270	4,155	1	,042	,577
	Age	,275	,039	48,711	1	,000	1,317
	Sum	-,165	,042	15,096	1	,000	,848
	Outcome	,183	,059	9,591	1	,002	1,200
	Income	,017	,048	,132	1	,716	1,017
	Education	,204	,036	31,924	1	,000	1,227
	Children	-,052	,036	2,118	1	,146	,949
	Estate	,503	,051	95,925	1	,000	1,654
	AlertsTotal	-,050	,014	12,015	1	,001	,951
	AlertsActive	-,158	,042	13,912	1	,000	,854
	Constant	,354	,106	11,286	1	,001	1,425
Step 5[a]	Sex(1)	-,321	,067	23,297	1	,000	,725
	workexpdum(1)	,210	,071	8,858	1	,003	1,234
	regions			22,964	4	,000	
	regions(1)	-,175	,069	6,510	1	,011	,839

	regions(2)	-,405	,157	6,657	1	,010	,667
	regions(3)	-,463	,123	14,113	1	,000	,629
	regions(4)	-,553	,270	4,185	1	,041	,575
	Age	,274	,039	48,631	1	,000	1,315
	Sum	-,160	,040	15,749	1	,000	,852
	Outcome	,192	,053	13,365	1	,000	1,212
	Education	,206	,036	33,234	1	,000	1,229
	Children	-,051	,035	2,029	1	,154	,951
	Estate	,504	,051	96,358	1	,000	1,655
	AlertsTotal	-,050	,014	11,895	1	,001	,952
	AlertsActive	-,158	,042	14,012	1	,000	,854
	Constant	,369	,097	14,352	1	,000	1,447
Step 6[a]	Sex(1)	-,298	,065	21,368	1	,000	,742
	workexpdum(1)	,209	,071	8,755	1	,003	1,232
	regions			24,997	4	,000	
	regions(1)	-,185	,068	7,369	1	,007	,831
	regions(2)	-,415	,157	6,990	1	,008	,660
	regions(3)	-,479	,123	15,226	1	,000	,619
	regions(4)	-,572	,269	4,514	1	,034	,564
	Age	,280	,039	51,464	1	,000	1,323
	Sum	-,163	,040	16,383	1	,000	,849
	Outcome	,182	,052	12,207	1	,000	1,200
	Education	,212	,036	35,602	1	,000	1,236
	Estate	,498	,051	94,986	1	,000	1,645
	AlertsTotal	-,051	,014	12,824	1	,000	,950
	AlertsActive	-,158	,042	13,915	1	,000	,854
	Constant	,330	,093	12,465	1	,000	1,390

Annex2: Sensitivity and Specificity

Coordinates of the Curve

Test Result Variable(s): Predicted probability

Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity
0,0000000	1,000	1,000
,1566467	1,000	,999
,1805043	1,000	,999
,2395063	,999	,997
,2469696	,999	,997
,3226410	,999	,991
,4042356	,991	,968

,4272555	,987	,948
,4719686	,970	,904
,4721130	,970	,904
,4958624	,958	,869
,5438548	,919	,787
,5441628	,919	,786
,5534491	,912	,773
,5535445	,911	,773
,5757204	,883	,721
,5757390	,883	,719
,5884874	,863	,684
,6489012	,758	,515
,6609727	,726	,489
,6995342	,640	,381
,7009959	,635	,368
,7010301	,635	,368
,7011545	,635	,368
,7012860	,634	,368
,7013105	,634	,368
,7014051	,633	,368
,7015209	,633	,367
,7015578	,632	,367
,7015797	,632	,367
,7017948	,632	,367
,7684262	,453	,205
,8173128	,303	,095
,8173699	,303	,094
,8174294	,303	,093
,9944335	,001	0,000
,9983410	,000	0,000
1,0000000	0,000	0,000

The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group.
a. The smallest cutoff value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test value plus 1. All the other cutoff values are the averages of two consecutive ordered observed test values.

Annex 3: Dependent and Independent Variables

Variables	Explanation
Status	Good clients(Y=1) who paid back their debts without facing any problem and bad clients(Y=0)
Sex	Female/Male; female is 0 and male is 1
Age	Age of customers (less than 31, 31-45, 46-60, more than 60)
Region	Countries such as Tartumaa, Põlvamaa … (15 countries)
Language	Mother tongue (Estonian, Russian)
Sum	Amount money is taken from bank (loan sum) (less than 250€,251€-500€,501€-750€ and more than 750€)
Period	Loan period in days
Income	Monthly income in EUR (less than 300€,301€-600€,601€-900€ and more than 900€)
Outcome	Monthly outcome in EUR (less than 300€,301€-600€,601€-900€ and more than 900€)
Family	Marriage Status (Single, Married)
Education	Education level (Basic education, High school, Vocational, University)
WorkExperience	Clients who has more than one year experience and less than one year
Children	Number of Children
Estate	Number of real estate units
AlertsTotal	Total number of payment problems
AlertsActive	Number of active payment problems
AlertsClosed	Number of closed payment problems

Non-exclusive licence to reproduce thesis and make thesis public

I, Tevfik Can Özay (date of birth: 21.02.1987),

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 Segmented Modelling”,supervised by Kalev Pärna,

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, **12.05.2016**