University of Tartu

Faculty of Social Sciences

School of Economics and Business Administration

Natia Kartsidze

Modeling short-term consumer loan defaults on example of European countries

Master's thesis

Supervisor: Oliver Lukason

Tartu 2019

Name and signature of supervisor…….

Allowed for defense on.....

(date)

I have written this master's thesis independently. All viewpoints of other authors, literary sources and data from elsewhere used for writing this paper have been referenced.

……..

(signature of author)

Table of contents

Abstract

The aim of the paper is to build a model distinguishing defaulting borrowers on short-term consumer loans. Customer characteristics selected considering available literature are fitted into a logistic model. The research is based on a sample dataset taken from the international finance company's database including 9 European countries with 6800 customers for each. The model performance and variable significance for different countries and when all countries' data is fitted into the model together is then analyzed. Results show that although the effect of some variables are consistent for all the countries, others vary in respect to sign and significance. The area under the curve for the model in all the cases is higher than 0.71 and classification accuracy higher than 65%.

1. Introduction

Short-term consumer loans are usually small loans with a term of around 30-60 days for the segment which is not well-served by traditional bank services. While bank procedures may take time, microcredit companies can provide the amount in minutes with just a couple of clicks online. Microfinance organizations were intended to become alternatives to the "loan-sharks"[1] that used to take advantage of the clients. The main purpose of such credit is to overcome unexpected cash shortage for emergency situations. The loans have high interest rates because of high risk caused by having no collateral[2] base.

Having a good customer base is key to being profitable in the business. As competition is quite intense and there are no significant switching barriers for clients, there is a high chance of losing a profitable customer if for some reason loan is not issued to him. As usual the companies have lower barriers when it comes to issuing loans to already existing customers, in order not to lose them to competitors. "At the customer management level, companies are striving ever harder to keep their existing clients by offering them additional products and enhanced services" (Siddiqi, 2006, p. 1). Companies should know their customers, their payment behavior to make better choices on the consequent loan issuance to minimize the credit losses. Moreover, time pressure forces institutions to develop automated processes for customer assessment and decision making. "Economic pressures resulting from increased demand for credit, allied with greater commercial competition and the emergence of new computer technology, have led to the development of sophisticated statistical models to aid the credit granting decision" (Hand, Henley, 1997, p. 523-524). To do that companies should understand what aspects of a customer trigger default.

Due to confidentiality of customer data, there are not many papers related to personal credit. Many previous papers were based on rather small sample size (Ozdemir, Boran, 2004; Abid, Masmoudi, Zouari-Ghorbel, 2016; Horkko, 2010; Lessmann, Baesens, Seow, Thomas, 2015) causing lack of credibility. Analyzing and predicting default appropriately

[1] A moneylender charging extremely high interest rates, usually with illegal conditions.
[2] Property or other asset pledged as security for repayment of a loan, which lender can seize in case the borrowers does not make promised payment.

is rather difficult, especially if there are data limitations. Most of the studies use U.S data and are also evolving for Asia, but a small amount of papers apply data from European countries (Horkko, 2010, p. 19).

The purpose of the paper is to build a model distinguishing defaulting borrower on short term, unsecured[3] loans. The model performance for different countries will be compared. The analysis will be based on real data from a multinational group focusing on microfinance services.

Most common methods used for such analyzes include neural networks, decision tree, logistic regression, discriminant analysis, Bayes classifier, k-nearest neighbor, support vector machines. There is no best model, and choice depends on the case, data and aim (Hand, Henley, 1997). The model chosen for the study is logistic regression based on the good fit for binomial outcomes and ease of interpretation.

In several studies logistic regression was compared to other methods and concluded to be better or as good as the other methods applied in the study (Kocenda, Vojtek 2011; Goncalves, Gouvea, 2007; Goriunov, Venzhyk, 2013; Abid, Masmoudi, Zouari-Ghorbel, 2016; Tsai, 2009). In the research similar to the paper Kocenda and Vojtek, 2011 concluded that logistic regression allows identification of the variables with the most discriminating power (distinguishability) in detecting default, which is more difficult or impossible in case of machine learning tools.

The dataset for the modeling includes customers from the following European countries: Bulgaria, Czech Republic, Denmark, Finland, Latvia, Lithuania, Poland, Spain and Sweden. These European countries were used as they are the ones where the company providing data operates. For each country a sample of 6800 customers, 3400 defaulted and 3400 not-defaulted was selected randomly from the microfinance organization's database and the model was fitted separately for each country and then all together, followed by comparison of the outcomes. For testing the model accuracy, a sample of 3000 customers, 1500 defaulted and 1500 non-defaulted, for each country was used different from the training dataset on which model was based.

[3] Not supported by a collateral, no asset pledged by a borrower to a lender

Variables selected are client age, gender, employment status (employee, unemployed, student, pensioner, board member, not specified, business owner, retiree, self-employed, other), accept news (whether the client agreed to receive news/marketing notifications from the company), master channel (the source for the application (direct, organic search, paid search, affiliate, offline)), previous loan months ago (how many months ago was a previous loan taken), number of paid loans, term, loan amount, maximum delay, average extension count, delay 5 count (count of events when the client had delay of 5 days or more on any of the previous loans).

The structure of the paper is as follows. The literature review will cover the description of consumer loan, methods used for default modeling, previous findings, definition of default and discussion about variables used for the modeling. Section three will be about model building, data, variable selection and fitting the sample into the model. The results will be presented in section four with evaluation of the model outcomes, country comparison and discussion on what happens if all countries' data is fitted together into one model altogether. Finally, everything will be summed up in the conclusion.

2. Literature review

There are not many papers on short-term consumer loans. Default behavior has been mostly studied in the area of mortgages[4] and corporate loans[5] (Galindo, Tamayo, 1997; Goriunov, Venzhyk, 2013; Zurada, Foster, Ward, Barker, 1999; Kofi, Portia, 2015; Lessmann, Baesens, Seow, Thomas, 2015). Due to limit of the studies in the sector of interest, references from the related fields are also considered to analyze the methods and the variables used. Appendix 5 summarizes papers in the field of interest or similar fields, the variables and method used, and the conclusions reached.

[4] Debt instrument which is secured by the collateral of specified real estate property. Borrower should pay back the loan and interest by pre-determined schedule.
[5] Loans made to businesses for a specific business purpose.

Consumer loan is an amount of money lent to individuals for personal purposes, overcoming unexpected money shortage, car repair, medical emergencies etc. Short-term consumer loans, also called personal loans, are part of consumer loans family, but unlike mortgages and other long-term loans they are unsecured and have term of around 30-60 days. Such service is usually provided by microfinance organizations rather than banks. As usual they use automated system to decide on loan issuance due to time pressure caused by competition, "Aggressive marketing efforts have resulted in deeper penetration of the risk pool of potential customers, and the need to process them rapidly and effectively has led to growing automation of the credit and insurance application and adjudication processes" (Siddiqi, 2006, p. 1). This requires understanding of customer characteristics having effect on default and developing a model discriminating between default and non-default.

In the review and comparison of classification methods applied in credit scoring, Louzada, Ara, Fernandes, 2016 state that the number of papers published each year from January 1992 to December 2015 ranges from 0 to 25 and are mostly from Belgium, United Kingdom, Taiwan, US, Chile and Brazil, restricting the study eligibility to journal papers in English, keywords 'credit scoring', 'machine learning', 'data mining', 'classification' or 'statistic'. From the methods used in the reviewed papers, logistic regression was most used (15.2%) in recent years and used as frequently as neural networks in the period after 2012.

Comparing neural networks and logit regression, it was concluded that neural networks are not superior to logistic regression for traditional dichotomous response variable. However, neural networks outperform for more complex financial distress variables with multiple response states. (Zurada, Foster, Ward, Barker, 1999)

Main classification methods in credit scoring are neural networks, support vector machine, linear regression, decision trees, logistic regression, fuzzy logic, genetic programming, discriminant analysis, Bayesian networks, hybrid methods, and ensemble methods (Louzada, Ara, Fernandes, 2016).

Logistic regression is a reliable model for analyzing data with dependent variable and several explanatory variables where dependent variable is binary. Although the model has

been criticized due to its inability to deal with multicollinearity[6], several studies support logistic regression mostly due to its predictive power (Goncalves, Gouvea, 2007; Abid, Masmoudi, Zouari-Ghorbel, 2016; Kocenda, Vojtek, 2011). The true advantage of the method is that it enables to clearly outline variables significantly distinguishing between defaulting and not defaulting borrowers (Kocenda, Vojtek, 2011). In the comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients it is stated that the main advantage of logistic regression is that it can produce a simple probabilistic formula of classification (Yeh, Lien, 2009).

"Classification methods which are easy to understand (such as regression, k-nearest neighbor and tree-based methods) are much more appealing, both to users and clients, than methods which are essentially black boxes (such as neural networks). They also permit more ready explanations of the sorts of reasons why the methods have reached their decisions." (Hand, Henley, 1996). In the same paper Hand, Henley, 1996 state that significant improvements are more likely the result from including new, more predictive characteristics rather than trying to experiment with more complex methods.

In the paper "Credit-scoring models in the credit-union environment using neural networks and genetic algorithms" predictive power of neural networks and genetic algorithms were compared to traditional techniques such as linear discriminant analysis and logistic regression. It was found that the traditional techniques compare very well with the more modest techniques (Desai, Conway, Crook, Overstreet, 1997).

The next two parts of the literature review will cover definition of default and variables used for the modeling.

2.1 Definition of default

Before developing a model, it is crucial to have 'Default' defined, i.e. what kind of behavior is considered to be default. It is a difficult step as there is no one universal definition and it depends on the researcher and the business insight. The concept differs

[6] Independent variables being linear functions of each other, high correlation between independent variables.

from company to company, depending on several aspects, such as client base and company policies. One may define default as 1-day delay or 30-days delay, or take write-off policy as baseline etc.

The main purpose of defining default is finding a point after which there is no chance that customer will pay. Existing portfolio of clients should be analyzed to find a suitable definition of default. This is usually done by analyzing delinquency for the portfolio, vintage or cohort analysis report. (Siddiqi, 2006)

The most frequent approach is using a delinquency of 90 days or more as default (Khandani, Kim, Lo, 2010; Hasan, 2016; Kocenda, Vojtek, 2011; Goriunov, Venzhyk, 2013; Charpignon, Horel, Tixier, 2014). Other definitions used in studies are listed below:

- Cases classified as default if, at any time in the last 48 months, the customer's most recent loan was charged off[7] or if the customer went bankrupt[8] (Desai, 1995).
- In three-way classification scheme, a case was classified as 'good' only if there were no payments that had been overdue for 31 days or more, 'poor' if the payment had ever been overdue for 60 days or more, and 'bad' if, at any time in the last 48 months, either the customer's most recent loan was charged off or the customer went bankrupt (Desai, Conway, Crook, Overstreet, 1997).
- Classified as defaulted in case of delinquency for 60 or more days and clients with a maximum delinquency of 20 days were considered as not-defaulted (Goncalves, Gouvea, 2007).
- Client being considered as defaulted if a contracted overdraft[9] is exceeded for more than 35 days during the period of 6 months (Sarlija, Bensic, Zekic-Susac, 2006).
- Default is if no payments were made in the last two months (Galindo, Tamayo, 1997).

The Basel Committee on Banking Supervision is an international committee which develops standards for banking regulation. They develop policy recommendations known

[7] Declared as unlikely to be collected.
[8] Legally declared as unable to pay debts.
[9] Available bank account balance gone below zero.

as Basel Accords. According to the Basel II requirement, receivables that are more than 90 days past due can be considered as defaulted, non-performing (Basel Committee on Banking Supervision, 2006). In the paper we will stick to Basel II Capital Accord definition of default as 90 days delinquency because as mentioned above, it is the definition preferred by the most authors and by the regulatory organization and it is reliable and easily interpretable.

2.2 Variables for modeling default

Table 1 below lists the most common variables used in the reviewed papers with their expected effect on default behavior based on the findings. Most commonly used variables are gender (Kocenda, Vojtek, 2011; Musto, Souleles, 2005; Thanh, Kleimeier, 2007), age (Thanh, Kleimeier, 2007; Musto, Souleles, 2005; Tsai, Lin, Cheng, Lin, 2009), loan amount (Ozdemir, Boran, 2004; Abid, Masmoudi, Zouari-Ghorbel, 2016; Kocenda, Vojtek, 201), employment (Kocenda, Vojtek, 2011; Abid, Masmoudi, Zouari-Ghorbel, 2016; Thanh, Kleimeier, 2007), loan term (Ozdemir, Boran, 2004; Thanh, Kleimeier, 2007; Kofi, Portia, 2015), education (Kocenda, Vojtek, 2011; Goriunov, Venzhyk, 2013; Thanh, Kleimeier, 2007; Tsai, Lin, Cheng, Lin, 2009; Constangioara, 2011), income (Charpignon, Horel, Tixie; Musto, Souleles, 2005; Thanh, Kleimeier, 2007; Ozdemir, Boran, 2004; Tsai, Lin, Cheng, Lin, 2009; Constangioara, 2011).

The amount of resources a client owns, the level of education, marital status, the purpose of the loan, and the years of having an account at the bank were found influential in a paper on the default predictors in retail credit scoring based on Czech banking data focusing mostly on socio-demographic characteristics, with logistic regression and classification and regression trees models applied (Kocenda, Vojtek, 2011). Ozdemir and Boran, 2004 concluded that from demographic characteristics, only occupation and residential status were related to default, while financial variables were found having significant influence. Research based on Ukrainian retail banking identified loan value, loan amount, term, contract type, gender, company type, work experience, family status, and credit history to be significant (Goriunov, Venzhyk, 2013). Thanh and Kleimeier,

2007 found time with bank, gender, number of loans, and loan duration to be the most valuable characteristics.

The paper introduces variables that were not found in the reviewed literature, such as accept news, master channel, average extension count, previous loan months ago, number of paid loans, considering their significance and novelty as Hand and Henley, 1996 state significant improvements are more likely result from including new characteristics.

Table 1. The most common variables in the reviewed papers and expected effect based on the effect found in the reviewed papers.

Variables	**Paper(s)**	**Expected effect**
Client age/date of birth	Abid, masmoudi, zouari-ghorbel, 2016; Goriunov, venzhyk, 2013; Goncalves, gouvea, 2007; Kocenda, vojtek, 2011; Tsai, lin, cheng, lin, 2009.	-
Gender	Goncalves, Gouvea, 2007; Kocenda, Vojtek, 2011; Tsai, Lin, Cheng, Lin, 2009; Ozdemir, Boran, 2004; Goriunov, Venzhyk, 2013.	+ (for males)
Employment status/occupation	Abid, masmoudi, zouari-ghorbel, 2016; Kocenda, vojtek, 2011; Tsai, lin, cheng, lin, 2009; Ozdemir, boran, 2004; Goriunov, venzhyk, 2013.	+/- (depends on categorization)
Education	Kocenda, Vojtek, 2011; Goriunov, Venzhyk, 2013; Thanh, Kleimeier, 2007; Tsai, Lin, Cheng, Lin, 2009; Constangioara, 2011.	-
Income	Charpignon, Horel, Tixie; Musto, Souleles, 2005; Thanh, Kleimeier, 2007; Ozdemir, Boran, 2004; Tsai, Lin, Cheng, Lin, 2009; Constangioara, 2011.	-
Marital status	Kocenda, Vojtek, 2011; Musto, Souleles, 2005; Thanh, Kleimeier, 2007; Ozdemir, Boran, 2004.	+ (for not married)
Dependants	Thanh, Kleimeier, 2007; Musto, Souleles, 2005; Charpignon, Horel, Tixier, 2014.	-
Years of employment	Kocenda, vojtek, 2011; Goriunov, venzhyk, 2013.	-
Purpose of loan	Kocenda, vojtek, 2011; Thanh, kleimeier, 2007; Kofi a. E., portia b., 2015.	+/- (depends on categorization)
Term	Goriunov, Venzhyk, 2013; Ozdemir, Boran, 2004; Thanh, Kleimeier, 2007.	+
Loan amount	Abid, masmoudi, zouari-ghorbel, 2016; Galindo, tamayo, 1997; Kocenda, vojtek, 2011; Ozdemir, boran, 2004; Goriunov, venzhyk, 2013.	+

Source: own elaboration based on the reviewed literature

Note: Expected effect shows whether default likelihood increases/decreases as a result of the increase in the variable value, + meaning increase, - decrease.

3. Method and data

Modeling default is based on fitting a model with several independent variables to describe binomial outcome of default or not default. However, which method is the best is very debatable. For the paper logistic regression is used due to facts mentioned in the literature review including the fact that compared to other more complex models e.g. neural networks, it allows outlining significant variables affecting default (Kocenda, Vojtek, 2011) and comparing them between different countries.

Logistic regression does not require many of the principle assumptions of linear regression models like normality of the error distribution or homoscedasticity[10] of the errors. Assumptions for logistic regression are no multicollinearity (independent variables being linear functions of each other, correlation between variables in the model which makes estimation insignificant and biased) and independence of the observations[11]. (Park, 2013) It has become standard method for the cases when dependent variable is binary (Hosmer, Lemenshow, 2000). Binary means taking only two values like true or false, male or female, default or no default.

It is important for the sample to be large enough to reduce the effect of possible multicollinearity, avoid overestimation of the effect measure and have statistically significant results. With the growth in the number of predictors, the sample size should also increase. Hosmer and Lemeshow, 2000 recommend sample sizes greater than 400.

Logit transformation is equal to the log of the odds. Odds are ratios of probabilities of an event happening, in our case default, to probabilities of the event not happening, here not default, as presented in the formula below. (Peng, Lee, Ingersoll, 2002)

$$odds = \frac{p}{1-p} \quad (1)$$

Here p is probability of event and 1-p probability of non-event.

[10] Constant variance of the residual or error term.
[11] Observations are independent if the occurrence of one provides no information about the occurrence of the other observation.

To linearize probability and have estimated probabilities in a range of 0 and 1, the natural logarithm from odds is taken as presented below (Park, 2013).

$$logit(y) = \ln(odds) = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta_i x_i \quad (2)$$

Here y is dependent variable, x_i independent variable, α represents the intercept and β_i the slope of the regression.

The odds ratio is a comparative measure of two odds relative to different events. The OR represents the odds that an outcome, in our case default or not default, will happen given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure. β (beta coefficient) shows the effect independent variable has on logged odds of the dependent, estimated measure of the rate of change of logit for a unit change in input variables. The exponential function of the regression coefficient β is the OR associated with a one unit increase in the independent variable. In case OR is 1 then exposure does not affect odds of outcome. OR being greated than one indicates exposure associated with higher odds of outcome and lower than one indicates exposure associated with lower odds of outcome. (Park, 2013)

For logistic regression there should not be missing values (Siddiqi, 2006), thus if there are any they should be handled. Possible ways for that are listed below.

- Exclude all data with missing values which might not be the best one in case the dataset is not big enough and removing missing value cases might end with very little data for analysis.
- Exclude variable with missing values from the model, with a drawback in case the variable has statistical significance.
- Include missing values as separate category.
- Use statistical values like mean or mode to replace missing value.

(Meeyai, 2016; Siddiqi, 2006)

Logistic regression is usually evaluated considering the following aspects:

- Overall model evaluation
- Evaluating each independent variable
- Evaluation of the predictability of the model

Overall model evaluation is about overall fit, relationship between the dependent variable and all the independent variables. There are several techniques to perform this. To assess it, the model without the independent variables can be compared with the model including the variables. The model without independent variables is considered as null model. If the model with independent variables performs better than the null model, then it is a better fit.

The overall fit can be tested by a likelihood test. It tests model with independent variables against null hypothesis, which is if all beta coefficients are equal to zero. Likelihood of the model is about likelihood of getting observation in case of no independent variable for null model and with independent variables for test model. Difference of those is $\chi 2$ statistic with the degrees of freedom same as the number of the independent variables in test model. As a result, if p-value is higher than the significance level (usually 0.05) then H_0 is rejected, meaning the test model is better than the null model.

Coefficients in a logistic regression show what is the effect on predicted log odds of having an outcome for a unit change in the independent variable. The likelihood ratio test and the Wald statistic are usually used for accessing the significance of the coefficients. The likelihood ratio test compares the probability of getting outcome when the parameter is zero with the probability of obtaining the data evaluated at the Maximum likelihood estimator of the parameter.

The classification table is used to check how well the model predicts outcomes. It summarizes sensitivity, specificity, false positives and false negatives. Sensitivity is the proportion of correctly classified events, specificity - the proportion of correctly classified nonevents, false positive - the proportion of observations misclassified as events, false negative - the proportion of observations misclassified as nonevents.

Receiver Operating Characteristic (ROC) curve is a useful tool for evaluating and comparing predictive models. It shows how well a predictive model can distinguish between the true positives and negatives. The curve plots sensitivity, the probability of predicting a real positive will be a positive, against 1-specificity, the probability of predicting a real negative will be a positive. The further the curve is from the diagonal line, the better the model is at discriminating between positives and negatives in general. Area under the curve (AUC) is the measure of how well the model performs in

distinguishing default and non-default correctly. The higher the AUC, the better the model.

The dataset for the modeling includes customers from the following European countries: Bulgaria, Czech Republic, Denmark, Finland, Latvia, Lithuania, Poland, Spain and Sweden, since those are where the company providing the data operates. For each country a sample of 6800 customers, 3400 defaulted, 3400 not-defaulted is selected randomly from the microfinance organization's database. For testing the model for each country 3000 customers, 1500 defaulted, 1500 not defaulted, is used different from the training dataset on which model is based.

Missing values are present for one variable - employment status, which is included in the model as a separate category as 'not specified'. Data for numeric variables is used as raw, continuous, without grouping into categories.

For selecting independent variables, available techniques include:

- Using all available variables.
- Forward selection means including first only the best characteristic according to individual predictive power and after adding the next best characteristics one by one until all significant characteristics are included and see how the model changes.
- Backward elimination is the opposite of the forward selection and starts with including all characteristics and removing insignificant ones one by one.
- Stepwise is a combination of the two above and involves adding and removing variables until the best version of the model is reached.

(Siddiqi, 2006)

For the case backward elimination method was applied. Initially 20 variables were fitted in the model. Due to correlation and insignificance several variables were removed. Removed variables are first loan months ago, count of events when delay days 10 were reached, count of events when delay days 30 were reached, count of events when delay days 60 were reached, count of event when delay days 90 were reached, source online or offline, last paid loan issued amount, number of paid loan in the last 1 year.

When selecting variables, there usually is temptation to use all that are available, however as usual problems arise such as multicollinearity, overfitting i.e. meaningless variables included (Park, 2013). 12 variables were selected (11 for countries other than Finland, Spain, Latvia, Czech Republic as employment status was not available for them). They are client age, gender, employment status, accept news, source (master) channel, previous loan months ago, paid loans, term, loan amount, maximum delay, average extension count, delay 5 count. Binary outcome for logistic regression event and non-event are defined as 1 when defaulted and 0 when not defaulted.

From the selected variables, gender (Kocenda, Vojtek, 2011; Musto, Souleles, 2005; Thanh, Kleimeier, 2007), age (Thanh, Kleimeier, 2007; Musto, Souleles, 2005; Tsai, Lin, Cheng, Lin, 2009), loan amount (Ozdemir, Boran, 2004; Abid, Masmoudi, Zouari-Ghorbel, 2016; Kocenda, Vojtek, 201), employment (Kocenda, Vojtek, 2011; Abid, Masmoudi, Zouari-Ghorbel, 2016; Thanh, Kleimeier, 2007), loan term (Ozdemir, Boran, 2004; Thanh, Kleimeier, 2007; Kofi A. E., Portia B., 2015) are commonly used in the similar papers. Some of the other commonly used variables, education (Kocenda, Vojtek, 2011; Goriunov, Venzhyk, 2013; Thanh, Kleimeier, 2007; Tsai, Lin, Cheng, Lin, 2009; Constangioara, 2011), income Charpignon, Horel, Tixie; Musto, Souleles, 2005; Thanh, Kleimeier, 2007; Ozdemir, Boran, 2004; Tsai, Lin, Cheng, Lin, 2009; Constangioara, 2011), marital status (Kocenda, Vojtek, 2011; Musto, Souleles, 2005; Thanh, Kleimeier, 2007; Ozdemir, Boran, 2004), dependants (Thanh, Kleimeier, 2007; Musto, Souleles, 2005; Charpignon, Horel, Tixier, 2014), years of employment (Kocenda, vojtek, 2011; Goriunov, venzhyk, 2013), purpose of loan (Kocenda, vojtek, 2011; Thanh, kleimeier, 2007; Kofi a. E., portia b., 2015), were not used in the paper due to data quality issues or the data provider not collecting such information.

Client age, previous loan months ago, delay 5 count, paid loans, term, loan amount, maximum delay, average extension count are continuous. Accept news is binary yes/no field. Source (master) channel, employment status and gender are also categorical variables. Categories in employment status are employee, unemployed, not specified, pensioner, self-employed, business owner, student, retiree, other. Categories in source (master) channel are: direct, affiliate, offline, organic search, paid search, other digital. The variables are summarized in the table 2 below with their type and definition.

Since employment status was not available for all the countries, first model without employment status was built to better compare model performance for all the countries and then model with employment status was run only for those countries where employment status was available. For both cases models with country included as variable were also built.

Table 2. Defining variables used in the paper.

Variable	Variable code name	Type	Definition
Client age	client_age	Socio-demographic	Age of the customer.
Accepting news	accept_news	Behavioral	Whether the client accepted to receive news/marketing notifications from the company. No as base category
Months since previous loan	previous_loan_months_ago	Behavioral	How long ago was the previous loan in months.
Paid loans	paid_loans	Behavioral	Number of paid loans.
Term	term	Behavioral	Term of the last applied loan.
Loan amount	loan_amount	Behavioral	Amount of the last applied loan.
Maximum delay	max_delay	Behavioral	The maximum delay the client has ever had.
Average extension count	avg_extension_count	Behavioral	Average of count of the extensions the client had on any previous loans.
Delay 5 count	delay_5	Behavioral	Count of events when the client had delay of 5 days on any of the previous loans.
Source channel	master_channel	Behavioral	What was the source for the application (direct, offline, organic search, paid search, affiliate, other digital). Affiliate as base category.
Employment status	employment_status	Socio-demographics	Customer's employment status (employee, unemployed, student, pensioner, board member, not specified, business owner, retiree, self-employed, other). Not specified as base category.
Gender	Gender	Socio-demographics	Gender of the client. Female as base category.

4. Results

Summary statistics for variables are presented in the appendices 3 and 4. Average age for the countries are mostly in a range 33-35, little higher for Spain, 40 and Latvia 38. Minimum age for all is between 18-20. Maximum age is 75-86 in most of the countries, except lower in Bulgaria - 65 and Lithuania - 70. Average months number from previous loans is 4-7 everywhere except higher number for Latvia - 12. Minimum months number is 0 everywhere while maximum varies being highest for Finland - 98, Latvia - 94, lowest in Czech Republic - 31 and others in a range of 48-59. Average number of paid loans for Finland is 10, others are in a range 3-7. Minimum number of paid loans is 0-1 for all, while maximum is relatively high for Finland - 100, Latvia - 69, others in a range 29-40. Average term is in a range of 27-29, maximum 30 everywhere except Finland, where it is 61, minimum is 1-2 in Finland, Denmark, Latvia, Poland, 5-7 in Czech Republic, Spain, Sweden and Bulgaria, 10 for Lithuania. Delay 5 count is 1-2 in all country datasets and minimum is 0, maximum is in 10-14 except Finland where it is 36. Average extension count is the range of 0-2 for all country datasets, with 0 as minimum and maximum in a range of 14-29. Minimum for maximum delay is 0 for all, average is relatively low for Bulgaria - 18, highest for Latvia - 197, in the range 61-77 for Finland, Denmark, Sweden, 35-56 for Czech Republic, Spain, Lithuania, Poland. Average amount is highest for Finland dataset - 1038, in 800-922 for Denmark and Sweden, in 280-550 for others.

Correlation lower than 0.5 is usually considered low, not risky (Hinkle, Wiersma, Jurs, 2003). The correlations between variables are below 0.5, so no multicollinearity risk is arising from this. Correlations for all countries' data together is presented in appendix 1. Multicollinearity can also be assessed by computing a variance inflation factor (VIF). It measures how much the variance of a regression coefficient is inflated because of multicollinearity in the model. Usually if VIF is smaller than ten then multicollinearity is of no risk, however, some authors consider VIF smaller than 5 as risk free (Alauddin, Nghiemb, 2010). By variable inflation check there is no risk for multicollinearity as for all the variables it is below 2.

Odd ratios are presented below in Table 3 and Table 4.

Table 3. Odds ratios when employment status is not included.

	CZ		FI		BG		ES		SE		LV		LT		DK		PL		ALL		ALL2	
Predictors	*Odds Ratios*	*p*	*Odds Ratios*	*p*	*Odds Ratios*	*p*	*Odds Ratios*	*p*	*Odds Ratios*	*p*	*Odds Ratios*	*p*	*Odds Ratios*	*p*	*Odds Ratios*	*p*	*Odds Ratios*	*p*	*Odds Ratios*	*p*	*Odds Ratios*	*p*
(Intercept)	0.44	**0.002**	0.25	**<0.001**	0.17	**<0.001**	0.17	**<0.001**	0.01	**<0.001**	0.04	**<0.001**	0.00	0.979	0.08	**<0.001**	0.80	0.555	0.17	**<0.001**	0.31	**<0.001**
client age	0.98	**<0.001**	0.98	**<0.001**	0.98	**<0.001**	0.99	**0.036**	0.99	**<0.001**	0.98	**<0.001**	0.97	**<0.001**	0.99	**<0.001**	0.98	**<0.001**	0.98	**<0.001**	0.98	**<0.001**
accept_newsY	0.50	**<0.001**	1.23	**0.001**	0.46	**<0.001**	0.13	**<0.001**	1.85	**<0.001**	1.28	**<0.001**	1.12	0.068	1.14	0.052	0.07	**<0.001**	0.62	**<0.001**	0.57	**<0.001**
previous loan months ago	0.96	**<0.001**	0.99	**0.005**	0.95	**<0.001**	0.96	**<0.001**	1.00	0.446	1.01	**<0.001**	0.97	**<0.001**	0.96	**<0.001**	1.00	0.832	0.99	**<0.001**	0.99	**<0.001**
paid loans	0.88	**<0.001**	0.98	**<0.001**	0.90	**<0.001**	0.91	**<0.001**	0.98	**<0.001**	0.96	**<0.001**	0.90	**<0.001**	0.92	**<0.001**	0.94	**<0.001**	0.95	**<0.001**	0.95	**<0.001**
term	1.04	**<0.001**	1.05	**<0.001**	1.07	**<0.001**	1.09	**<0.001**	1.09	**<0.001**	1.08	**<0.001**	1.08	**<0.001**	1.07	**<0.001**	1.08	**<0.001**	1.08	**<0.001**	1.07	**<0.001**
loan amount	1.00	**<0.001**	1.00	**<0.001**	1.00	**<0.001**	1.00	**<0.001**	1.00	**<0.001**	1.01	**<0.001**	1.00	**<0.001**	1.00	**<0.001**	1.00	**<0.001**	1.00	**<0.001**	1.00	**<0.001**
max delay	1.01	**<0.001**	1.00	**<0.001**	1.00	**<0.001**	1.00	**<0.001**	1.00	0.781	1.00	**<0.001**	1.00	**<0.001**	1.00	**<0.001**	1.00	**<0.001**	1.00	**<0.001**	1.00	**<0.001**
avg extension count	1.02	0.212	0.90	**<0.001**	1.03	0.121	0.99	0.327	1.06	**<0.001**	0.98	**0.044**	0.95	**<0.001**	0.98	0.372	0.98	0.283	1.00	0.269	0.99	**0.004**
delay 5	1.12	**<0.001**	1.08	**<0.001**	1.09	**0.006**	1.04	0.236	1.12	**<0.001**	1.09	**0.001**	1.01	0.814	1.06	**0.030**	1.15	**<0.001**	1.08	**<0.001**	1.12	**<0.001**
master_channelDirect	0.76	0.185	0.55	**<0.001**	1.15	0.586	0.59	0.098	0.90	0.663	0.56	0.090	1736434.03	0.982	0.71	0.101	0.44	**0.004**	0.86	0.051	0.73	**<0.001**
master_channelOffline	0.86	0.457	0.20	**0.025**	1.99	**0.009**	0.49	**0.011**	1.70	**0.035**	0.72	0.330	2922572.92	0.981			0.48	**0.010**	1.08	0.278	0.80	**0.004**
master_channelOrganic Search	0.74	0.162	0.64	**0.005**	1.08	0.766	0.47	**0.009**	1.18	0.528	0.57	0.091	2.52	0.999	0.75	0.176	0.37	**0.001**	0.86	0.050	0.76	**<0.001**
master_channelOther Digital	0.68	0.058	0.44	**<0.001**	1.52	0.099	0.56	**0.030**	2.41	**<0.001**	0.50	**0.037**	2167612.02	0.981	0.77	0.180	0.33	**<0.001**	0.94	0.408	0.83	**0.012**
master_channelPaid Search	0.74	0.199	0.78	0.391	0.48	0.409	0.26	**0.027**	0.66	0.099	0.00	0.936	31.16	0.997	0.84	0.421	0.51	0.659	0.69	**<0.001**	0.69	**<0.001**
gendermale	0.97	0.628	1.10	0.137	1.14	**0.016**	1.01	0.860	1.18	**0.003**	1.36	**<0.001**	1.19	**0.001**	1.28	**<0.001**	0.93	0.233	1.12	**<0.001**	1.12	**<0.001**
countryCzech_rep																					0.66	**<0.001**
countryDenmark																					0.39	**<0.001**
countryFinland																					0.25	**<0.001**
countryLatvia																					1.07	0.075
countryLithuania																					0.78	**<0.001**
countryPoland																					0.57	**<0.001**
countrySpain																					0.59	**<0.001**
countrySweden																					0.30	**<0.001**
Observations	6800		6800		6800		6800		6800		6800		6800		6800		6800		61200		61200	
Cox & Snell's R^2 / Nagelkerke's R^2	0.192 / 0.256		0.309 / 0.411		0.151 / 0.201		0.313 / 0.417		0.235 / 0.314		0.217 / 0.290		0.167 / 0.223		0.206 / 0.275		0.348 / 0.464		0.154 / 0.205		0.178 / 0.237	

Note: All is all countries together without employment status and All2 is the same plus country as variable.

Table 4. Odd ratios only for countries with employment status.

	CZ		FI		ES		LV		ALL		ALL2	
Predictors	*Odds Ratios*	*p*	*Odds Ratios*	*p*	*Odds Ratios*	*p*	*Odds Ratios*	*p*	*Odds Ratios*	*p*	*Odds Ratios*	*p*
(Intercept)	0.46	**0.009**	0.42	**<0.001**	0.15	**<0.001**	0.04	**<0.001**	0.39	**<0.001**	0.33	**<0.001**
client age	0.98	**<0.001**	0.97	**<0.001**	0.99	0.085	0.99	**<0.001**	0.98	**<0.001**	0.98	**<0.001**
accept_newsY	0.50	**<0.001**	1.23	**0.001**	0.13	**<0.001**	1.29	**<0.001**	0.56	**<0.001**	0.54	**<0.001**
previous loan months ago	0.96	**<0.001**	0.99	**0.018**	0.96	**<0.001**	1.01	**<0.001**	0.99	**<0.001**	1.00	**0.012**
paid loans	0.88	**<0.001**	0.98	**<0.001**	0.92	**<0.001**	0.96	**<0.001**	0.95	**<0.001**	0.96	**<0.001**
term	1.04	**<0.001**	1.05	**<0.001**	1.09	**<0.001**	1.08	**<0.001**	1.07	**<0.001**	1.07	**<0.001**
loan amount	1.00	**<0.001**	1.00	**<0.001**	1.00	**<0.001**	1.01	**<0.001**	1.00	**<0.001**	1.00	**<0.001**
max delay	1.01	**<0.001**	1.00	**<0.001**	1.00	**<0.001**	1.00	**<0.001**	1.00	**<0.001**	1.00	**<0.001**
avg extension count	1.03	0.179	0.90	**<0.001**	0.99	0.374	0.98	**0.041**	1.00	0.959	0.98	**0.007**
delay 5	1.12	**<0.001**	1.09	**<0.001**	1.04	0.215	1.09	**0.001**	1.09	**<0.001**	1.14	**<0.001**
master_channelDirect	0.76	0.191	0.56	**<0.001**	0.59	0.094	0.57	0.091	0.70	**0.001**	0.63	**<0.001**
master_channelOffline	0.84	0.392	0.20	**0.029**	0.49	**0.011**	0.73	0.353	0.73	**0.003**	0.56	**<0.001**
master_channelOrganic Search	0.74	0.157	0.67	**0.011**	0.47	**0.009**	0.57	0.094	0.67	**<0.001**	0.63	**<0.001**
master_channelOther Digital	0.68	0.055	0.44	**<0.001**	0.55	**0.029**	0.50	**0.037**	0.67	**<0.001**	0.57	**<0.001**
master_channelPaid Search	0.73	0.180	0.83	0.541	0.26	**0.029**	0.00	0.936	0.77	0.086	0.63	**0.002**
employment_statusEMPLOYEE	0.99	0.955	0.54	**<0.001**	1.02	0.872	0.94	0.649	0.54	**<0.001**	0.89	0.062
employment_statusOTHER	1.10	0.484							0.82	**<0.001**	1.13	0.130
employment_statusPENSIONER	0.81	0.860	1.19	0.312	1.06	0.744	0.91	0.863	0.66	**<0.001**	1.21	**0.041**
employment_statusRETIREE	1.24	0.254							0.72	**0.009**	1.07	0.647
employment_statusSELF_EMPLOYED	0.72	0.056			1.04	0.816			0.60	**<0.001**	0.83	0.051
employment_statusSTUDENT	0.77	0.249	0.82	0.236			1.19	0.886	0.52	**<0.001**	1.17	0.182
employment_statusUNEMPLOYED	0.74	0.226	1.02	0.893	1.39	**0.025**	0.57	0.216	0.77	**<0.001**	1.30	**0.001**
gendermale	0.99	0.907	1.13	**0.049**	1.02	0.802	1.36	**<0.001**	1.15	**<0.001**	1.12	**<0.001**
employment_statusBUSINESS_OWNER			0.28	**<0.001**					0.15	**<0.001**	0.32	**<0.001**
employment_statusMEMBER OF THE BOARD							0.19	0.160	0.33	0.332	0.32	0.319
countryFinland											0.37	**<0.001**
countryLatvia											1.54	**<0.001**
countrySpain											0.92	0.086
Observations	6800		6800		6800		6800		27200		27200	
Cox & Snell's R^2 / Nagelkerke's R^2	0.194 / 0.259		0.321 / 0.427		0.314 / 0.419		0.218 / 0.290		0.189 / 0.251		0.204 / 0.272	

Note: All is all countries together with employment status and All2 is the same plus country as variable.

When odds ratio is bigger than 1 then effect of the variable is positive on default, thus it increases chance of default and when it is smaller than 1 then negative. Below behavior for each variable is summarized and compared between countries and all country together models by sign of the effect and significance.

1. Loan amount and term are the only variables significant in each of the countries and all together models with positive effect on likelihood of default behavior.
2. Number of paid loans has negative effect and is significant in all the models.
3. Accept news is significant in every model except for Lithuania, however the sign is not the same everywhere. It has negative effect for Poland, Czech Republic, Spain, Bulgaria, all countries together models and positive everywhere else.
4. Age is significant for all models except Spain with employment status model and it has negative effect everywhere.
5. Previous loan months ago is significant with negative effect in all except Latvia where it has positive effect for both with and without employment status and for Poland and Sweden it is not significant.
6. Delay 5 count is significant with positive effect everywhere except Lithuania and Spain.
7. Maximum delay is significant everywhere except Sweden and with positive effect.
8. Average extension count appeared significant for Finland, Latvia, Sweden and all country models where country is included as variable.
9. Gender is significant for Latvia, Sweden, Denmark, Lithuania, Finland with employment status, Bulgaria, all country model without employment status and all country models where country is included as variable and effect is positive meaning for males there is higher chance for the default behavior.
10. Master channel does not have any noticeable pattern. It is significant for some models with different signs and not significant for others.
11. For the countries where employment status was available and all country models with employment status, in Spain compared to not specified, unemployed is significant with plus effect on default, while other categories appear to be not significant, in Finland business owner and employee are significant with negative effect, in all countries with employment status model all except board member are

significant with negative effect compared to not specified and for all countries with employment status model with country included as variable unemployed is significant with positive effect on default.

Compared with the expected behavior based on the reviewed papers summarized in the table 1, behavior of gender, age, amount and term are in accordance to expected effect.

Area under the curve for the model for each of the countries and all together models are summarized in the table 5 and ROC curves are presented in Appendix 2. AUC is usually interpreted as follows:

- AUC=0.5 No discrimination
- 0.6≥AUC>0.5 Poor discrimination
- 0.7≥AUC>0.6 Acceptable discrimination
- 0.8≥AUC>0.7 Excellent discrimination
- AUC>0.9 Outstanding discrimination (Yang, Berdine, 2017; Mandrekar, 2010, Hosmer, Lemeshow, 2000)

As presented in the table 5, AUC for every country and all together models are in the range 0.7-0.85. The best result observed was for Poland 0.85, which was beyond expectation as models with employment status would be more likely to perform better and for Poland the variable was not available. For all country models, including country as variable slightly improves model performance. For countries for which employment status was available, comparing models with and without employment status, for Finland and Czech Republic with employment status model is slightly better while for Spain and Latvia it is slightly worse, thus modeling without employment status included does not make the model much worse overall.

Sensitivity and specificity were calculated using two different cut-offs: the default 0.5 and by the optimal cutoff (when probability that gives minimum misclassification error is chosen) and the results are presented in the tables 6 and 7 below. Confusion matrices are presented in appendices 6 and 7. The classification results are higher than 65% for all countries and all countries models, which is better than classification only by chance.

Table 5. Area under the curve.

Country	**AUC with employment status**	**AUC without employment**
Finland	0.7753496	0.7752533
Bulgaria	0.7168522	
Spain	0.7980676	0.7982618
Sweden	0.7843649	
Latvia	0.7774013	0.7774947
Lithuania	0.7447204	
Poland	0.8535896	
Denmark	0.7515851	
Czech Republic	0.7284196	0.7248262
All countries		0.7200953
Only the countries with employment status	0.7206090	
All countries, country as variable		0.7356081
Only the countries with employment status, country as variable	0.7365657	

Table 6. Classification accuracy when optimal cut-off is used.

Country	**Cut-off**	**Specificity**	**Sensitivity**	**Overall**
Finland(with employment status)	0.45928	0.776	0.645	0.7105
Finland(without employment status)	0.33912	0.725	0.687	0.706
Bulgaria	0.48281	0.587	0.763	0.675
Spain(with employment status)	0.58773	0.777	0.701	0.739
Spain(without employment status)	0.58956	0.775	0.701	0.738
Sweden	0.45701	0.676	0.794	0.735
Latvia(with employment status)	0.46541	0.623	0.809	0.716
Latvia(without employment status)	0.46537	0.624	0.808	0.716
Lithuania	0.46826	0.590	0.789	0.6895
Poland	0.61090	0.827	0.755	0.791
Denmark	0.46912	0.634	0.743	0.6885
Czech Republic(with employment status)	0.45870	0.585	0.766	0.6755
Czech Republic(without employment status)	0.42926	0.531	0.807	0.669
All countries without employment status	0.47972	0.613	0.711	0.662
Only countries with employment status	0.47985	0.546	0.788	0.667
All countries without employment status, country as variable	0.46997	0.604	0.759	0.682
Only countries with employment status, country as variable	0.48995	0.663	0.698	0.681

Table 7. Classification accuracy when 0.5 is taken as cut-off.

Country	**Specificity**	**Sensitivity**	**Overall**
Finland(with employment status)	0.813	0.598	0.7055
Finland(without employment status)	0.869	0.498	0.6835
Bulgaria	0.611	0.721	0.666
Spain(with employment status)	0.723	0.745	0.734
Spain(without employment status)	0.724	0.745	0.7345
Sweden	0.715	0.729	0.722
Latvia(with employment status)	0.663	0.758	0.7105
Latvia(without employment status)	0.664	0.756	0.71
Lithuania	0.633	0.729	0.681
Poland	0.784	0.787	0.7855
Denmark	0.691	0.681	0.686
Czech Republic(with employment status)	0.641	0.695	0.668
Czech Republic(without employment status)	0.633	0.689	0.661
All countries without employment status	0.657	0.660	0.6585
Only countries with employment status	0.588	0.741	0.6645
All countries without employment status, country as variable	0.660	0.696	0.678
Only countries with employment status, country as variable	0.679	0.68	0.6795

Pseudo R squared does not work for logistic regression the same way as R squared works for linear regression. Maximum likelihood estimates for logistic regression are not calculated to minimize variance. Pseudo R squared can be used for comparison of the same type, same data, same outcome models in model building stage to compare models (Hosmer, Lemeshow, 2000) rather than evaluating single model. The higher the number the better.

Cox & Snell R^2 and Nagelkerke's R^2 are presented in table 3 and 4 above. The R^2 results do not look very good, but as usual they tend to be low in practice (Goriunov, Venzhyk, 2013; Musto, Souleles, 2005; Hosmer, Lemeshow, 2000). They vary for countries, some being relatively better than others, having relatively higher numbers for Spain, Poland and Finland.

5. Conclusion

In the paper logistic model was constructed to model default based on the sample for 9 European countries with 6800 customers for each. 12 variables were selected each of them showing significance in all or some of the countries and combined models. In addition to separate country models, the model with all the countries included together was also run.

Correlations between variables for each of the model were below critical 0.5 threshold and variance inflation factor below 5 meaning multicollinearity should not be of concern. Some of the variables show the same sign effect and significance in all the models, specifically term, amount and number of paid loans, while others change sign like accept news or become non-significant for some of the countries like gender, delay count 5, max delay, previous loan months ago. Variable being significant in least of the models was average extension count. From employment status significant categories included unemployed and pensioner with positive effect, business owner and employee with negative.

AUC for all the models is good enough, but not excellent, falling in 0.7-0.85 interval with the highest for Poland, 0.85, followed by Spain with almost 0.8. Overall classification accuracies are higher than 65% for all the models.

From the all country together models, the ones where country was included as variable performed slightly better than without the country variable. For the countries for which employment status was available, the model with employment status performed slightly better considering AUC and classification results. However, even if employment status is not available almost equally good model can be built.

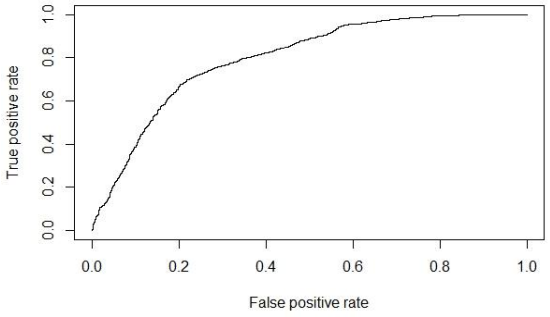
Overall, the research showed that when the same model is used for different countries, some variables can behave the same, while other can show completely opposite behavior. The most important variables were amount, term, paid loans and maximum delay and those are the ones that showed consistency through the models. Amount and term were among the most commonly used variables in the reviewed literature together with gender, income, employment, age. In line with the findings of Ozdemir and Boran, 2004 financial variables were found to have significant influence. However, compared to Kocenda,

Vojtek, 2011 the model did not include as many socio-demographic variables due to data issues, which might be considered as limitations for the paper and a point to consider for future enhancements. Additionally, it would be interesting to develop the model by machine learning techniques, e.g. neural networks, and see how it will compare to the logistic model.

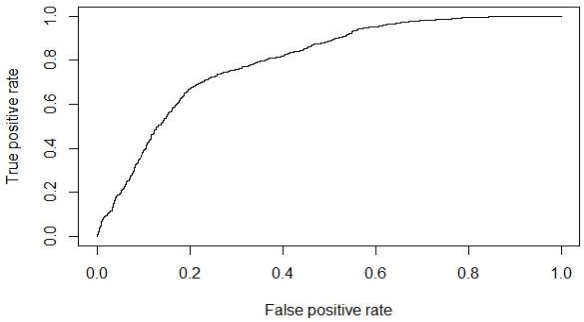
Appendices

Appendix 1. Receiver operating characteristic curve

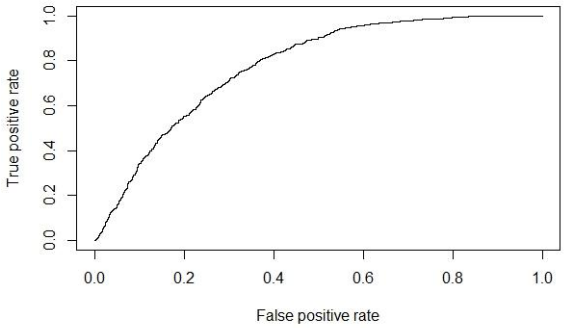
1)Spain with employment status



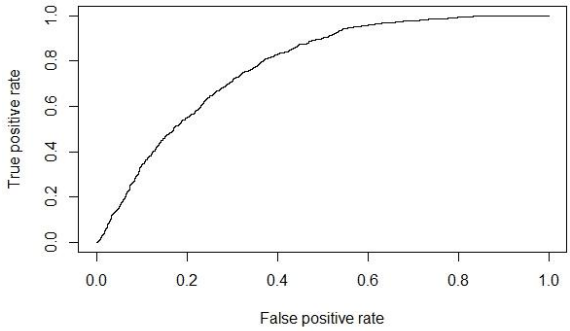
2)Spain without employment status



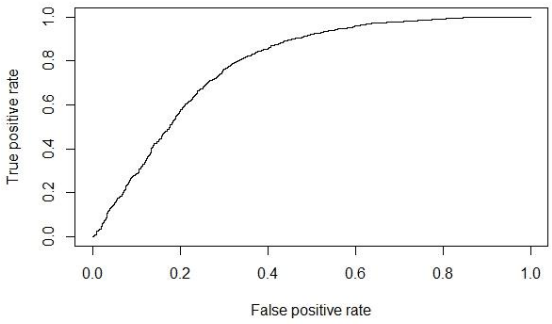
3)Latvia with employment status



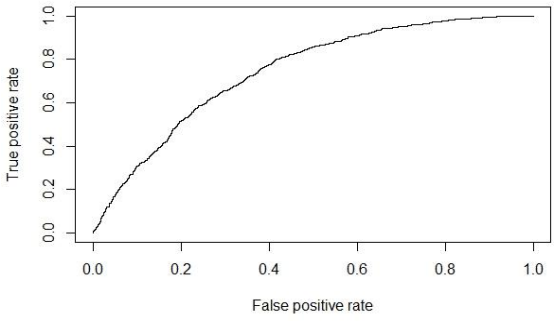
4)Latvia without employment status



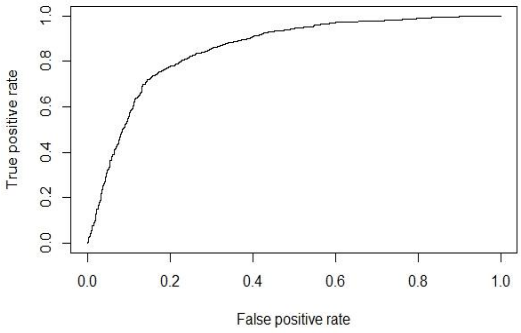
5)Sweden



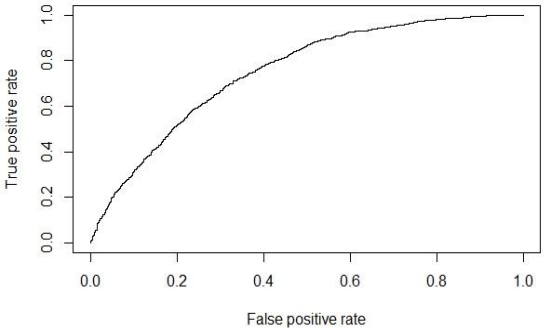
6)Lithuania



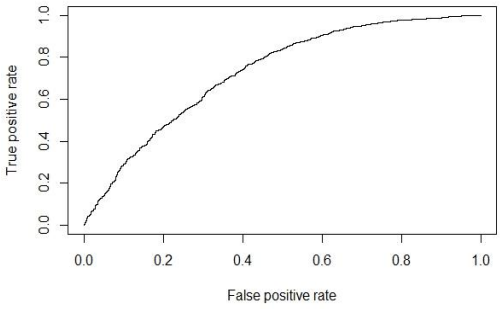
7)Poland



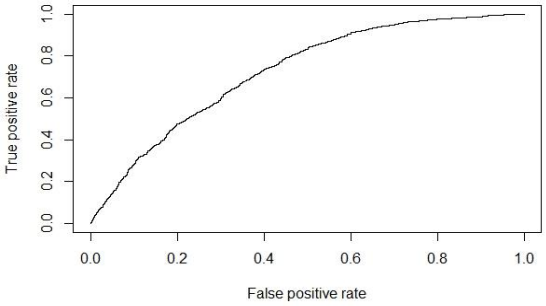
8)Denmark



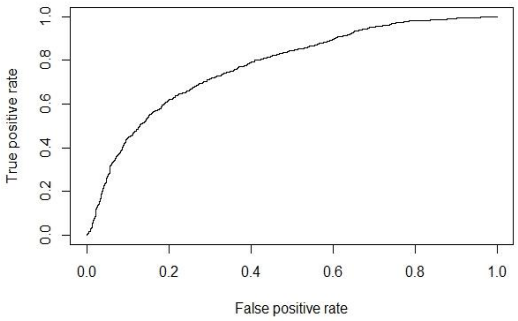
9)Czech Rep. with employment status



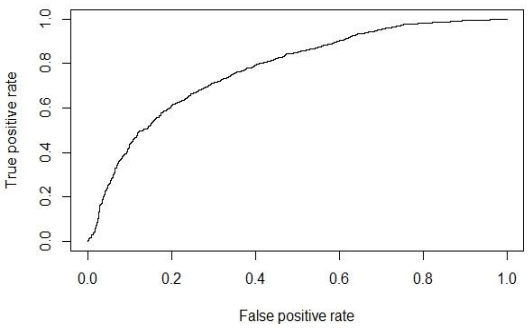
10)Czech Rep. without employment status



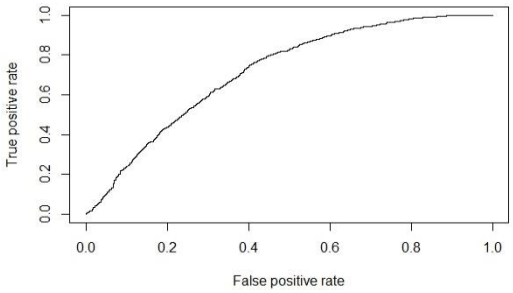
11)Finland with employment status



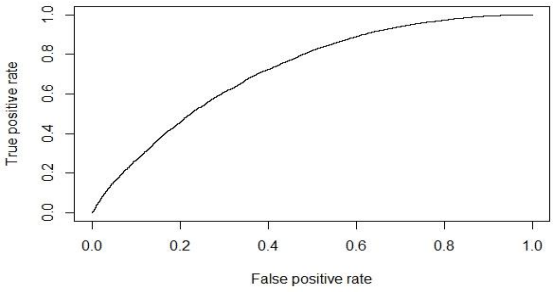
12)Finland without employment status



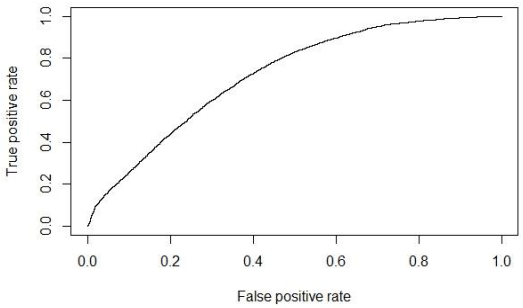
13)Bulgaria



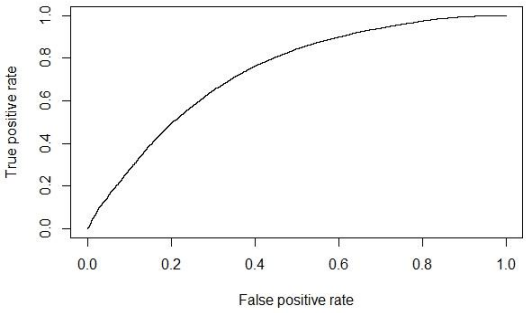
14)All countries together without employment status



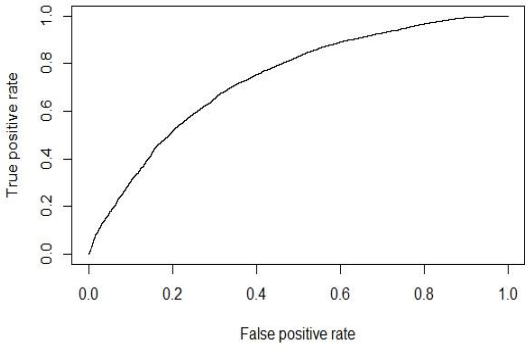
15)All countries with employment status



16)All without employment status with country as variable



17)All countries with employment status with country as variable



Appendix 2. Correlation between the variables.

	Client age	**Previous loan Months ago**	**Paid loans**	**Term**	**Loan amount**	**Max. delay**	**Avg. extension count**	**Delay 5**
Client age	1	0.068602	0.054628	0.051903	0.093304	-0.01281	0.090259	-0.03911
Previous loan months ago	0.068602	1	-0.09881	0.03385	-0.0147	0.352389	0.342753	0.078052
Paid loans	0.054628	-0.09881	1	-0.1627	0.154181	0.017808	-0.14378	0.314578
Term	0.051903	0.03385	-0.1627	1	0.216326	0.03946	0.076863	0.017135
Loan amount	0.093304	-0.0147	0.154181	0.216326	1	0.068383	-0.00604	0.231637
Max. delay	-0.01281	0.352389	0.017808	0.03946	0.068383	1	0.151565	0.242102
Avg. extension count	0.090259	0.342753	-0.14378	0.076863	-0.00604	0.151565	1	0.05477
Delay 5	-0.03911	0.078052	0.314578	0.017135	0.231637	0.242102	0.05477	1

Appendix 3. Summary statistics.

1) Finland

	Client_age	Previous_loan_months_ago	Paid_loans	Term	Loan_amount	Max_delay	Avg_extension_count	Delay_5
Mean	35.33	6.72	10.26	27.59	1037.97	75.41	0.39	2.37
Stdev	13.23	10.77	10.88	11.03	792.32	117.25	1.18	3.50
Median	31.00	3.00	7.00	30.00	800.00	17.00	0.00	1.00
Minimum	18.00	0.00	0.00	2.00	10.00	0.00	0.00	0.00
Maximum	80.00	98.00	100.00	61.00	3950.00	960.00	20.00	36.00
Observations	6800	6800	6800	6800	6800	6800	6800	6800

2) Czech Republic

	Client_age	Previous_loan_months_ago	Paid_loans	Term	Loan_amount	Max_delay	Avg_extension_count	Delay_5
Mean	34.18	4.41	3.38	27.87	439.12	35.27	0.75	0.87
Stdev	12.91	4.81	3.32	5.47	225.69	63.17	1.68	1.30
Median	30.00	3.00	2.00	30.00	463.62	4.00	0.00	0.00
Minimum	19.00	0.00	0.00	7.00	38.64	0.00	0.00	0.00
Maximum	75.00	31.00	29.00	30.00	1220.88	458.00	19.00	12.00
Observations	6800	6800	6800	6800	6800	6800	6800	6800

3) Denmark

	Client_age	Previous_loan_months_ago	Paid_loans	Term	Loan_amount	Max_delay	Avg_extension_count	Delay_5
Mean	34.76	5.99	4.32	27.51	809.06	60.59	0.68	0.83
Stdev	12.38	8.00	4.88	5.89	454.75	129.96	1.53	1.30
Median	31.00	3.00	3.00	30.00	806.22	4.00	0.00	0.00
Minimum	20.00	0.00	0.00	1.00	40.31	0.00	0.00	0.00
Maximum	80.00	55.00	39.00	30.00	3090.52	910.00	15.00	10.00
Observations	6800	6800	6800	6800	6800	6800	6800	6800

4) Spain

	Client_age	Previous_loan_months_ago	Paid_loans	Term	Loan_amount	Max_delay	Avg_extension_count	Delay_5
Mean	39.68	4.86	5.20	27.35	536.13	39.64	1.13	0.63
Stdev	11.29	7.00	5.77	5.98	240.13	96.23	2.65	1.09
Median	38.00	2.00	3.00	30.00	500.00	3.00	0.00	0.00
Minimum	18.00	0.00	1.00	7.00	50.00	0.00	0.00	0.00
Maximum	86.00	49.00	39.99	30.00	1150	713.00	27.00	10.00
Observations	6800	6800	6800	6800	6800	6800	6800	6800

5) Sweden

	Client_age	Previous_loan_months_ago	Paid_loans	Term	Loan_amount	Max_delay	Avg_extension_count	Delay_5
Mean	35.93	6.13	6.33	28.69	922.04	76.90	1.17	1.30
Stdev	12.37	7.51	6.77	4.40	564.37	277.13	2.33	1.86
Median	33.00	3.00	4.00	30.00	896.33	8.00	0.00	1.00
Minimum	18.00	0.00	1.00	5.00	31.27	0.00	0.00	0.00
Maximum	80.00	49.00	39.00	30.00	3298.72	2862.00	19.00	14.00
Observations	6800	6800	6800	6800	6800	6800	6800	6800

6) Bulgaria

	Client_age	Previous_loan_months_ago	Paid_loans	Term	Loan_amount	Max_delay	Avg_extension_count	Delay_5
Mean	33.44	3.99	3.29	28.44	281.01	17.77	0.57	0.57
Stdev	11.40	5.20	3.38	4.85	163.78	47.75	1.48	0.99
Median	30.00	2.00	2.00	30.00	255.23	2.00	0.00	0.00
Minimum	19.00	0.00	1.00	5.00	25.52	0.00	0.00	0.00
Maximum	65.00	48.00	28.00	30.00	2404.27	499.00	18.00	10.00
Observations	6800	6800	6800	6800	6800	6800	6800	6800

7) Latvia

	Client_age	Previous_loan_months_ago	Paid_loans	Term	Loan_amount	Max_delay	Avg_extension_count	Delay_5
Mean	37.77	11.94	7.35	28.29	321.79	197.21	1.86	0.93
Stdev	13.96	16.48	9.21	5.16	159.95	462.08	3.31	1.28
Median	34.00	4.00	4.00	30.00	375.00	6.00	0.00	1.00
Minimum	18.00	0.00	0.00	1.00	30.00	0.00	0.00	0.00
Maximum	75.00	94.00	69.00	30.00	1083.42	2998.00	29.00	10.00
Observations	6800	6800	6800	6800	6800	6800	6800	6800

8) Lithuania

	Client_age	Previous_loan_months_ago	Paid_loans	Term	Loan_amount	Max_delay	Avg_extension_count	Delay_5
Mean	30.58	6.86	4.56	28.78	367.47	55.57	1.57	0.65
Stdev	11.72	9.05	5.28	4.04	218.57	133.58	2.81	1.02
Median	26.00	3.00	3.00	30.00	362.03	3.00	0.00	0.00
Minimum	18.00	0.00	1.00	10.00	28.96	0.00	0.00	0.00
Maximum	70.00	59.00	39.00	30.00	1782.00	972.00	25.00	10.00
Observations	6800	6800	6800	6800	6800	6800	6800	6800

9) Poland

	Client_age	Previous_loan_months_ago	Paid_loans	Term	Loan_amount	Max_delay	Avg_extension_count	Delay_5
Mean	34.28	5.01	5.36	28.52	509.53	38.69	0.72	0.87
Stdev	12.69	7.32	5.84	5.03	368.94	96.84	1.70	1.38
Median	30.00	2.00	3.00	30.00	442.27	4.00	0.00	0.00
Minimum	19.00	0.00	1.00	1.00	23.28	0.00	0.00	0.00
Maximum	78.00	50.00	39.00	30.00	1792.35	797.00	14.00	10.00
Observations	6800	6800	6800	6800	6800	6800	6800	6800

10) All countries together

	Client_age	Previous_loan_months_ago	Paid_loans	Term	Loan_amount	Max_delay	Avg_extension_count	Delay_5
Mean	35.10	6.21	5.56	28.12	580.46	66.34	0.98	1.00
Stdev	12.70	9.36	6.90	6.11	483.62	207.08	2.23	1.77
Median	32.00	3.00	3.00	30.00	434.43	4.00	0.00	0.00
Minimum	18.00	0.00	0.00	1.00	10.00	0.00	0.00	0.00
Maximum	86.00	98.00	100.00	61.00	3950	2998	29.00	36.00
Observations	61200	61200	61200	61200	61200	61200	61200	61200

11) Only countries with employment status together

	Client_age	Previous_loan_months_ago	Paid_loans	Term	Loan_amount	Max_delay	Avg_extension_count	Delay_5
Mean	36.74	6.98	6.55	27.77	583.75	86.88	1.03	1.20
Stdev	13.06	11.13	8.27	7.32	514.78	253.82	2.42	2.16
Median	34.00	3.00	3.00	30	425.00	5.00	0.00	0.00
Minimum	18.00	0.00	0.00	1.00	10.00	0.00	0.00	0.00
Maximum	86.00	98.00	100.00	61.00	3950	2998	29.00	36.00
Observations	27200	27200	27200	27200	27200	27200	27200	27200

Appendix 4. Categorical variables summary.

1) Finland

Accept_news (2 distinct values).

Value	Y	N
Frequency	3491	3309

Master_channel (6 distinct values.)

Value	Affiliate	Direct	Offline	Organic search	Other digital	Paid search
Frequency	304	1500	15	2051	2841	89

Employment_status (6 distinct values).

Value	Unemployed	Business_owner	Employee	Not specified	Pensioner	Student
Frequency	604	268	4536	425	510	457

Gender (2 distinct values).

Value	Female	Male
Frequency	2848	3952

2) Czech Republic

Accept_news (2 distinct values).

Value	Y	N
Frequency	4113	2687

Master_channel(6 distinct values).

Value	Affiliate	Direct	Offline	Organic search	Other digital	Paid search
Frequency	138	895	2268	698	2530	271

Employment_status (8 distinct values).

Value	Unemployed	Self_employed	Employee	Not specified	Pensioner	Student	Other	Retiree
Frequency	114	377	3276	328	5	145	2230	325

Gender (2 distinct values).

Value	Female	Male
Frequency	3075	3725

3) Denmark

Accept_news (2 distinct values).

Value	Y	N
Frequency	5210	1590

Master_channel (5 distinct values).

Value	Affiliate	Direct	Organic search	Other digital	Paid search
Frequency	154	893	541	4460	752

Gender (2 distinct values).

Value	Female	Male
Frequency	2798	4002

4) Spain

Accept_news (2 distinct values).

Value	Y	N
Frequency	4253	2547

Master_channel (6 distinct values).

Value	Affiliate	Direct	Offline	Organic search	Other digital	Paid search
Frequency	91	248	832	613	4994	22

Employment_status (5 distinct values).

Value	Unemployed	Employee	Not specified	Pensioner	Self_employed
Frequency	1249	4126	375	589	461

Gender (2 distinct values).

Value	Female	Male
Frequency	2451	4349

5) Sweden

Accept_news (2 distinct values).

Value	Y	N
Frequency	1938	4862

Master_channel (6 distinct values).

Value	Affiliate	Direct	Offline	Organic search	Other digital	Paid search
Frequency	152	404	220	198	5562	264

Gender (2 distinct values).

Value	Female	Male
Frequency	2971	3829

6) Bulgaria

Accept_news (2 distinct values).

Value	Y	N
Frequency	1069	5731

Master_channel (6 distinct values).

Value	Affiliate	Direct	Offline	Organic search	Other digital	Paid search
Frequency	80	895	1040	740	4037	8

Gender (2 distinct values).

Value	Female	Male
Frequency	3039	3761

7) Latvia

Accept_news (2 distinct values).

Value	Y	N
Frequency	5101	1699

Master_channel (6 distinct values).

Value	Affiliate	Direct	Offline	Organic search	Other digital	Paid search
Frequency	56	928	1937	833	3042	4

Employment_status (6 distinct values).

Value	Unemployed	Member of board	Employee	Not specified	Pensioner	Student
Frequency	26	5	252	6492	22	3

Gender (2 distinct values).

Value	Female	Male
Frequency	3195	3605

8) Lithuania

Accept_news (2 distinct values).

Value	Y	N
Frequency	4973	1827

Master_channel (6 distinct values).

Value	Affiliate	Direct	Offline	Organic search	Other digital	Paid search
Frequency	2	21	1452	29	5295	1

Gender (2 distinct values).

Value	Female	Male
Frequency	3125	3675

9) Poland

Accept_news (2 distinct values).

Value	Y	N
Frequency	3261	3539

Master_channel (6 distinct values).

Value	Affiliate	Direct	Offline	Organic search	Other digital	Paid search
Frequency	88	992	1127	918	3672	3

Gender (2 distinct values).

Value	Female	Male
Frequency	3196	3604

10) All countries together without employment status

Accept_news (2 distinct values).

Value	Y	N
Frequency	39289	21911

Master_channel (6 distinct values).

Value	Affiliate	Direct	Offline	Organic search	Other digital	Paid search
Frequency	1065	6776	8891	6621	36433	1414

Gender (2 distinct values).

Value	Female	Male
Frequency	26698	34502

11) Only countries with employment status together

Accept_news (2 distinct values).

Value	Y	N
Frequency	15252	11948

Master_channel (6 distinct values).

Value	Affiliate	Direct	Offline	Organic search	Other digital	Paid search
Frequency	589	3571	5052	4195	13407	386

Employment_status (10 distinct values).

Value	Unemp.	Board member	Employee	Not specified	Pensioner	Student	Business owner	Other	Retiree	Self-employed
Freq.	1993	5	12190	7620	1126	605	268	2230	325	838

Gender (2 distinct values).

Value	Female	Male
Frequency	15631	11569

Appendix 5. Summary of the previous studies.

	Study	Methods	Dataset	Variables	Results
1	Default Predictors in Retail Credit Scoring: Evidence from Czech Banking Data (Kocenda, Vojtek, 2011).	Logistic regression, Classification and Regression Trees (CART)	3,403 individual clients	Education, marital status, years of employment, sector of employment, sex, date of birth, type of employment, number of employments, employment position, credit ratios, region, own resources, amount of loan, purpose of loan, length of the relationship, date of account opening, deposit behaviors, loan protection, type of product, number of co-signers, date of loan.	Logistic regression allows identification of the variables with the most discriminating power in detecting default. The methods have similar performance and do not significantly differ in predictive power. The amount of resources a client owns, the level of education, marital status, the purpose of the loan, and the years of having an account at the bank were found influential. Also, even if the amount of resources a client owns is not included in the model, performance still is only slightly worse than when it is included.
2	Credit Risk Analysis Applying Logistic Regression and Neural Networks Models (Goncalves, Gouvea, 2007).	Logistic Regression, Neural Networks	10,000 good contracts and 10,000 bad	Gender, marital status, home telephone, commercial telephone, time in the present job, quantity of loan parts, first acquisition, time in the present home, loan part value, type of credit, age, range of home ZIP code, range commercial ZIP Code, profession code, percent rate of part/salary, percent rate of loan/salary.	Results obtained by the logistic regression and neural network models were evaluated as good and very similar, although logistic regression was slightly better.
3	Credit Risk Assessment using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications (Galindo, Tamayo, 1997).	Probit, Decision-Tree CART model, Neural Networks, K-Nearest Neighbors	Mortgage loan dataset, 4000 customers	Credit amount, unpaid balance, overdue balance, debt, guarantee, guarantee1, guarantee2, interest, residential, acquisition, construction, liquidity, ten variables month1-month10 showing payment history.	CART decision-tree model was concluded to provide the best estimation for default.
4	Loan Default Prediction in Ukrainian Retail Banking (Goriunov, Venzhyk, 2013).	Logistic regression, Neural Networks	1348 car loans, 1821 mortgage loans	Loan/value, payment/income, cycle, currency of a loan, loan amount, loan term, interest rate, loan type, age, gender, resident, company, occupation, experience, military, education, family, recommendations, home phone, mobile, bank cards, card type, accounts, history, real estate, car, assets.	As the result Neural Networks was found only slightly better than Logistic regression, but it is usually the case when response variable has more than 2 outcomes. Out of the variables for car loans loan value, loan amount, term, contract type, gender, company type, work experience, family status, and credit history appeared to be significant. For mortgages the same variables in addition with price-to-income, loan term, real estate, and assets.
5	Neural Networks Versus Logit Regression Models For Predicting Financial Distress Response Variables (Zurada, Foster, Ward, Barker, 1999).	Logistic regression, Neural networks	204 firms		Neural networks did not perform better than Logistic regression for dichotomous response variable but performed better for multi-state response variable.

6	The Consumer Loan's Payment Default Predictive Model: An Application In A Tunisian Commercial Bank (Abid, Masmoudi, Zouari-Ghorbel, 2016).	Logistic regression, discriminant analysis	633 consumer loans	Age, loan amount, outstanding credit, occupational category.	By the results, logistic regression model outperformed discriminant analysis with logistical regression having an overall accuracy of 89%, and discriminant analysis - 68.49%. The results of the logistic regression showed that three factors were relevant in predicting default: loan amount having negative impact on default, outstanding loan and socio-professional category with positive effect.
7	Prediction of consumer credit risk (Charpignon, Horel, Tixier, 2014).	Logistic regression, Classification and Regression Trees (CART), Random Forests, Gradient Boosting Trees (GBT).	100,000 consumers, only 6% defaulted	Age of the borrower, number of dependents in family, monthly income, monthly expenditures divided by monthly gross income, total balance on credit cards divided by the sum of credit limits, number of open loans and lines of credit, number of mortgage and real estate loans, number of times the borrower has been 30-59 days past due but no worse in the last two years, number of times the borrower has been 60-89 days past due but no worse in the last two years, number of times the borrower has been 90 days or more past due.	Authors interpreted results as in favor of GBT model.
8	Development of a Credit Scoring Model for Retail Loan Granting Financial Institutions from Frontier Markets (Hasan, 2016).	Logistic regression	3000 "bad" and 3000 "good" retail loans	New/existing clients, Requested amount, Payment-to-Income ratio (PTI), Loan type.	Even with scarce data a model can be developed that can help assessing clients.
9	A Portfolio View Of Consumer Credit (Musto, Souleles, 2005).	Asset-Pricing Theory	U.S. credit bureau, Experian, data was used including approximately 100,000 randomly sampled consumers	All trades, non-revolving, revolving, card limits, credit score, beta, income, age, marriage, sex, kids, adults, home ownership, Business ownership, demographics, unemployment, insure, divorce.	Covariance risk tends to be higher for younger and single consumers, lower-income consumers, renting home, with higher rates of divorce and lower rates of health-insurance coverage and the amount of credit obtained by consumers significantly decreases with their covariance risk.
10	Credit Scoring for Vietnam's Retail Banking Market: Implementation and Implications for Transactional versus Relationship Lending (Thanh, Kleimeier, 2007).		56037 loans from Vietnam's commercial banks	Income, education, occupation, employer type, time with employer, age, gender, region, time at the present address, residential status, marital status, number of dependents, home phone, mobile phone, loan purpose, collateral type, collateral value, loan duration, time with the bank, number of loans, current account, savings account.	The most valuable variables they found were time with bank, gender, number of loans, and loan duration.

11	Examining Credit Default Risk: An Empirical Study on Consumer Credit Clients (Ozdemir, Boran, 2004).	Logistic regression	500 individuals with consumer credit	Credit category, interest rate, sex, age, marital status, income, loan size, maturity, residential status and occupation.	From demographic characteristics, only occupation and residential status were found relating to default, while financial variables were found having significant influence.
12	The consumer loan default predicting model – An application of DEA–DA and neural network (Tsai, Lin, Cheng, Lin, 2009).	Logistic Regression, Neural Networks, Discriminant Analysis, DEA–discriminant analysis (DEA–DA).	A sample from a certain financial institution in Taiwan with 1877 consumer loans	Gender, age, monthly income, education, occupation, power, retention.	They found that the predictive efficiency with all these four methods was more than 75%.
13	Consumer Credit Scoring (Constangioara, 2011).	Logistic regression, neural networks, decision tree and bagging estimations.	A Hungarian dataset of 5060 observations of existing accounts of loans for personal needs	Age, education, employment sector, family income, free income.	Bagging, logit and neural networks were found superior to traditional tree estimation.
14	Estimation of Default Probability on a Consumption Credit Portfolio of a Cabo Verde Bank by means of Logistic Regression (Fernandes, Esquıvel, Guerreiro, Xufre, Martins, 2012).	Logistic Regression	7183 applications	Number of monthly payments, the amount of monthly payments, age, occupation, employer, gender, agency, qualifications, monthly payment, nominal rate and type of guarantee presented by the client.	The variables found as the most important financial and behavioral characteristics of default behavior were employer, agency, benefits paid, provision of value, nominal rate, amount, occupation, age, gender and type of guarantee.
15	Determinants of business loan default in Ghana (Kofi, Portia, 2015).	Logistic regression	224 business customers of a bank in Ghana	Ownership type, owner's collateral, extra source of income, business age, business size, relationship with lender, multiple borrowing, business location, diversion of loan purpose, purpose of the loan, age of the loan/term, repayment plan/schedule, loan price, underfunding, delays in loan processing.	The study found that owner's extra income, multiple borrowing, diversion of loan purpose, loan price, loan purpose, loan age, repayment plan and underfunding to be significant in determining the probability of business loan default.
16	Credit-scoring models in the credit-onion environment using neural networks and genetic algorithms (Desai, Conway, Crook, Overstreet, 1997).	Neural networks, genetic algorithms, linear discriminant analysis, logistic regression.	962 observations for credit union L, 918 observations for credit union M, and 853 observations for credit union N	Major credit cards, owns home, income, bureau rating, job time, dependants, number of inquiries, trade age, trade line 75% full, payments as a proportion of income, deliquent accounts in the past 12 months, total debt as proportion of income, age, number of years at the current address, number of open accounts, number of previous loans.	Study concludes that traditional techniques compare very well with the two new techniques studied. Neural networks performed somewhat better than the rest of the methods for classifying the most difficult group, namely poor loans.

Appendix 6. Confusion matrices when optimal cut-off is used.

FINLAND (WITH EMPLOYMENT STATUS)		
	0	**1**
0	1164	532
1	336	968
SPAIN (WITH EMPLOYMEMT STATUS)		
	0	**1**
0	1166	449
1	334	1051
LATVIA(WITH EMPLOYEMENT STATUS)		
	0	**1**
0	934	287
1	566	1213
CZECH REP. (WITH EMPLOYMENT STATUS)		
	0	**1**
0	878	351
1	622	1149
DENMARK		
	0	**1**
0	951	385
1	549	1115
SWEDEN		
	0	**1**
0	1014	309
1	486	1191
POLAND		
	0	**1**
0	1240	368
1	260	1132
ALL COUNTRIES WITHOUT EMPLOYMENT STATUS		
	0	**1**
0	8277	3889
1	5223	9611

FINLAND (WITHOUT EMPLOYMENT STATUS)		
	0	**1**
0	1088	469
1	412	1031
SPAIN (WITHOUT EMPLOYMEMT STATUS)		
	0	**1**
0	1162	448
1	338	1052
LATVIA(WITHOUT EMPLOYEMENT STATUS)		
	0	**1**
0	936	288
1	564	1212
CZECH REP. (WITHOUT EMPLOYMENT STATUS)		
	0	**1**
0	796	290
1	704	1210
BULGARIA		
	0	**1**
0	880	356
1	620	1144
LITHUANIA		
	0	**1**
0	884	316
1	616	1184
ONLY COUNTRIES WITH EMPLOYMENT STATUS		
	0	**1**
0	3273	1271
1	2727	4729
ALL COUNTRIES WITHOUT EMPLOYMENT STATUS, COUNTRY AS VARIABLE		
	0	**1**
0	8156	3258
1	5344	10242
ONLY COUNTRIES WITH EMPLOYMENT STATUS, COUNTRY AS VARIABLE		
	0	**1**
0	3977	1814
1	2023	4186

Note: 1 denotes default, 0 not default

Appendix 7. Confusion matrices when 0.5 is used as cut-off.

FINLAND (WITH EMPLOYMENT STATUS)		
	0	**1**
0	1219	603
1	281	897
SPAIN (WITH EMPLOYMEMT STATUS)		
	0	**1**
0	1085	382
1	415	1118
LATVIA(WITH EMPLOYEMENT STATUS)		
	0	**1**
0	994	363
1	506	1137
CZECH REP. (WITH EMPLOYMENT STATUS)		
	0	**1**
0	962	458
1	538	1042
DENMARK		
	0	**1**
0	1036	478
1	464	1022
SWEDEN		
	0	**1**
0	1073	406
1	427	1094
POLAND		
	0	**1**
0	1176	320
1	324	1180
ALL COUNTRIES WITHOUT EMPLOYMENT STATUS		
	0	**1**
0	8870	4592
1	4630	8908

FINLAND (WITHOUT EMPLOYMENT STATUS)		
	0	**1**
0	1303	753
1	197	747
SPAIN (WITHOUT EMPLOYMEMT STATUS)		
	0	**1**
0	1086	382
1	414	1118
LATVIA(WITHOUT EMPLOYEMENT STATUS)		
	0	**1**
0	996	366
1	504	1134
CZECH REP. (WITHOUT EMPLOYMENT STATUS)		
	0	**1**
0	950	466
1	550	1034
BULGARIA		
	0	**1**
0	917	419
1	583	1081
LITHUANIA		
	0	**1**
0	950	407
1	550	1093
ONLY COUNTRIES WITH EMPLOYMENT STATUS		
	0	**1**
0	3530	1553
1	2470	4447
ALL COUNTRIES WITHOUT EMPLOYMENT STATUS, COUNTRY AS VARIABLE		
	0	**1**
0	8911	4102
1	4589	9398
ONLY COUNTRIES WITH EMPLOYMENT STATUS, COUNTRY AS VARIABLE		
	0	**1**
0	4071	1918
1	1929	4082

Note: 1 denotes default, 0 not default

References

1) Abid, L., Masmoudi, A., Zouari-Ghorbel, S., 2016. The Consumer Loan's Payment Default Predictive Model: An Application in A Tunisian Commercial Bank, Asian Economic and Financial Review, 6(1), 27-42.
2) Alauddin, M., Nghiemb, H. S., 2010. Do Instructional Attributes pose Multicollinearity Problems? An Empirical Exploration, Economic Analysis & Policy, 40(3), 351-361.
3) Basel Committee on Banking Supervision, 2006. International convergence of capital measurement and capital standards.
4) Charpignon, M., Horel, E., Tixier, F., 2014. Prediction of consumer credit risk, CS 229 projects, Stanford University.
5) Constangioara, A., 2011. Consumer Credit Scoring, Romanian Journal of Economic Forecasting, 0(3), 162-177.
6) Desai, V. S., Conway, D. G., Crook, J. N., Overstreet G. A., 1997. Credit-scoring models in the credit-onion environment using neural networks and genetic algorithms, IMA Journal of Management Mathematics, 8(4), 323-346.
7) Desai, V. S., Crook, J. N., Overstreet, G. A., 1995. A comparison of neural networks and linear scoring models in the credit union environment, European Journal of Operational Research, 95(1), 24-37.
8) Fernandes, J. M., Esquivel, M. L., Guerreiro, G., Xufre, P., Martins M. O. Estimation of default probability on a consumption credit portfolio of a Cabo Verde bank by means of logistic regression, paper presented at Joint Meeting of y-BIS and jSPE, 2012.
9) Galindo, J., Tamayo, P., 1997. Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications, 15(1), 107-143.
10) Goncalves, E. B., Gouvea, M. A. Credit Risk Analysis Applying Logistic Regression and Neural Networks Models, paper presented at 17th Annual Conference of POMS, 2007.
11) Goriunov, D., Venzhyk, K. Loan Default Prediction in Ukrainian Retail Banking. Working paper number 13/07E, 2013.
12) Hand, D. J., Henley, W. E., 1996. Statistical classification methods in consumer credit scoring: a review, 160(3), 523-541.
13) Hasan, K. R., 2016. Development of a Credit Scoring Model for Retail Loan Granting Financial Institutions from Frontier Markets, International Journal of Business and Economics Research, 5(5), 135-142.
14) Hinkle, E., Wiersma, W., Jurs, S., 2003. Applied Statistics for the Behavioral Sciences, 5th edition.
15) Hosmer, D. W., Lemeshow, S., 2000. Applied Logistic regression, 2nd edition.
16) Horkko, M., 2010. The Determinants of Default in Consumer Credit Market (Masters dissertation).
17) Khandani, A. E., Kim, A. Lo, A. W., 2010. Consumer Credit Risk Models via Machine-Learning Algorithms, Journal of Banking & Finance, 34(11), 2767-2787.

18) Kofi, A. E., Portia, B., 2015. Determinants of business loan default in Ghana, Junior Scientific Researcher, 1(1), 10-26.
19) Lessmann, S., Baesens, B., Seow, H., Thomas, L. C., 2015. Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research, European Journal of Operational Research, 247(1), 124-136.
20) Louzada, F., Ara, A., Fernandes, G. B., 2016. Classification methods applied to credit scoring: A systematic review and overall comparison, Surveys in operations research and management science, 21(2), 117-134.
21) Mandrekar, J., 2010. Receiver operating characteristic curve in diagnostic test assessment, 5(9), 1315-1316.
22) Meeyai, S., 2016. Logistic regression with missing data: a comparison of handling methods, and effects of percent missing values, Journal of Traffic and Logistics Engineering, 4(2), 128-134.
23) Musto, D. K., Souleles, N. S. A portfolio view of consumer credit, NBER working paper series, working paper number 11735, 2005.
24) Ozdemir, O., Boran, L. Examining Credit Default Risk: An Empirical Study on Consumer Credit Clients, working paper number 2004/20, 2014.
25) Park, H., 2013. An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain, Applied Mathematical Sciences, 12(4), 185-204.
26) Peng, J. C., Lee, K. L., Ingersoll, G. M., 2002. An introduction to logistic regression analysis and reporting, The Journal of Educational Research, 96(1), 3-14.
27) Sarlija, N., Bensic, M., Zekic-Susac, M., 2006. Modeling customer revolving credit scoring using logistic regression, survival analysis and neural networks, proceedings of the 7th WSEAS international conference on neural networks, 164-169.
28) Siddiqi, N., 2006. Credit risk scorecards: developing and implementing intelligent credit scoring.
29) Thanh, D. T., Kleimeier, S., 2007. Credit scoring for Vietnam's retail banking market: implementation and implications for transactional versus relationship lending, 16(5), 471-495.
30) Tsai, M., Lin, S., Cheng, C., Lin, Y., 2009. The consumer loan default predicting model – An application of DEA–DA and neural network, expert systems with applications, 36(9), 11682-11690.
31) Vojtek, M., Kocenda, E. Default predictors in retail credit scoring: evidence from Czech banking data, William Davidson Institute, working paper number 1015, 2011.
32) Yang, S., Berdine, G., 2017. The receiver operating characteristic (ROC) curve, The Southwest Respiratory and Critical Care Chronicles, 5(19), 34–36.
33) Yeh, I., Lien, C., 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients, Expert Systems with Applications, 36(2), 2473-2480.
34) Zurada, J. M., Foster, B. P., Ward, J. T., Barker, R. M., 1999. Neural networks versus logit regression models for predicting financial distress response variables, The journal of applied business research, 15(1), 21-30.

Non-exclusive licence to reproduce thesis and make thesis public

I, Natia Kartsidze,

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,

Modeling short-term consumer loan defaults on example of European countries,

supervised by Oliver Lukason,

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, **14.01.2019**