University of Tartu

Faculty of Science and Technology

Institute of Mathematics and Statistics

Suleyman Ahmadov

# Weight of Evidence Methodology in Logistic Regression with Application in Credit Scoring

Actuarial and Financial Engineering

Master's Thesis (30 ECTS)

Supervisor: Prof. emer. Kalev Pärna

Tartu 2023

# Weight of Evidence Methodology in Logistic Regression with Application in Credit Scoring

Master Thesis

Suleyman Ahmadov

## Abstract

A popular technique for improving the prognostication capability of logistic regression models is the Weight of Evidence (WoE) methodology. This study examines the use of WoE methodology in credit scoring with the goal of evaluating how well it enhances the precision and understandability of credit risk assessment models. An extensive dataset containing data on the demographics, finances and credit history of credit seekers is used in this study. The study examines how the WoE methodology is used to logistic regression models and evaluates how well it performs in comparison to more conventional methods. The study uses a comparative analysis framework to compare the effectiveness of the WoE and conventional logistic regression models using a variety of evaluation metrics, such as accuracy, precision, recall, and the area under the receiver operating characteristics curve (AUC-ROC). Furthermore, by looking at the weights given to various variables and their importance in the model, the interpretability of the WoE techniques is evaluated.

By proving the effectiveness of the WoE methodology in credit scoring models based on logistic regression, the findings of this thesis make a contribution to the field of credit risk assessment. The findings provide financial institutions and lenders with insightful information that helps them decide whether to implement the WoE approach for credit risk assessment. This study uses logistic regression and WoE approach to create comprehensive and understandable credit scoring models, with ultimate goal of enhancing credit risk management procedures.

**CERCS research specialization**: P160 Statistics, programming, financial and actuarial mathematics.

**Keywords**: Weight of Evidence, WoE, logistic regression, credit scoring.

# Tõendite kaalukuse metoodika logistilises regressioonis koos rakendusega krediidiskooringus

Magistritöö
Suleyman Ahmadov

**Lühikokkuvõte**

Populaarne tehnika logistiliste regressioonimudelite prognoosimisvõime parandamiseks on tõendite kaalukuse (WoE) metoodika. Antud töös uuritakse WoE metoodika kasutamist krediidiskooringus eesmärgiga hinnata, kui hästi see suurendab krediidiriski hindamise mudelite täpsust ja arusaadavust. Lõputöös kasutatakse ulatuslikku andmestikku, mis sisaldab infot krediiditaotlejate demograafia, rahanduse ja krediidiajaloo kohta. Uuringus näidatakse, kuidas WoE metoodikat kasutada logistiliste regressioonimudelite korral, ja hinnatakse, kui hästi see toimib võrreldes tavapärasemate meetoditega. Et võrrelda WoE ja tavapäraste logistiliste regressioonimudelite tõhusust, kasutatakse erinevaid hindamismõõdikuid nagu täpsus, tagasikutsumine ja ROC-kõvera alune pindala (AUC-ROC). Peale selle, vaadeldes erinevatele muutujatele antud kaalusid ja nende tähtsust mudelis, hinnatakse WoE tehnika tõlgendatavust.

Näidates WoE metoodika efektiivsust logistilisel regressioonil põhinevates krediidiskooringu mudelites, annavad käesoleva lõputöö tulemused kasuliku panuse krediidiriski hindamise valdkonda. Töös pakutakse finantsasutustele ja laenuandjatele põhjalikku teavet, mis aitab neil otsustada, kas on mõistlik rakendada krediidiriski hindamisel WoE-meetodit. Uuringus kasutatud logistilise regressiooni ja tõendite kaalukuse lähenemisviisid aitavad luua terviklikke ja arusaadavaid krediidiriski hindamise mudeleid, mille lõppeesmärk on tõhustada krediidiriski juhtimise protseduure.

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

**Märksõnad**: Tõendite kaalukus, WoE, logistiline regressioon, krediidiskooring.

# Contents

# 1. Introduction

## 1.1 Credit scoring

In the modern day, financial institutions, the most prominent of which are commercial banks, use statistical analysis to establish the creditworthiness of a physical or legal persons. Credit and Scoring should be separated into their two individual parts according to Anderson's (2007, p.7) definition of credit scoring. First off, the word "credit" itself simply means "buy now, pay later". The Latin word "credo", which meaning "I believe" or "I trust", is the source of the word. The other meaning of "scoring" is "the use of a numerical tool to evaluate order cases according to some real or perceived quality, for being able to differentiate between them, and to guarantee consistent, objective decisions. Because of this, scores could be displayed as "numbers" to represent a single quality or as "grades" that might be shown as "letters" or "labels" to represent one or more qualities (Anderson, 2007, p.9). A risk management tool that evaluates the creditworthiness of a party is known as a credit scoring model, i.e., the ability to repay the loan, of a loan applicant by estimating his/her probability of default based on historical data.
Lenders frequently make two different kinds of decision: first, whether to approve credit for a new application and second, how to handle an existing application, including whether or not to increase the credit limits. (Thomas, Crook & Edelman, 2002, p.7).

Although financial institutions have existed since at least 2000 BC, the history of credit scoring is relatively new. A credit applicant's information is gathered by banks and/or financial organizations, and this data is utilized to create a numerical score for each application (Hand & Jacka, 1998, p.18). Credit scoring methods have recently been broadened to encompass new applications in many industries. Additionally, the concept of lowering a customer's likelihood of defaulting – which foretells customer risk – is a new role for credit scoring, which can support aid in maximizing the expected profit from that customer for financial institutions, particularly banks. The usage of credit scoring was growing steadily at the dawn of the twenty-first century, especially as a result of the amazing technological advancements that brought about more sophisticated methodologies and evaluation standards. Nonetheless, as calculating abilities of technologies increase, more new methods emerge.

**Example of Credit Scoring type: FICO scoring method**

The "FICO" is one of the most widely spread credit scoring techniques used in many countries, one of those countries that adapts the Fico-score is United States.
The range of the borrowing party's credit worthiness on the FICO scale is from 300 up to 850.

| Credit category | Credit Score | Percentage |
|---|---|---|
| Exceptional credit | 800-850 | ≈ 20% |
| Very Good credit | 740-799 | ≈ 25% |
| Good credit | 670-739 | ≈ 21% |
| Fair credit | 580-669 | ≈ 18% |
| Poor credit | <580 | ≈ 16% |

https://www.myfico.com/credit-education/what-is-a-fico-score

The FICO credit scoring formula is obtained after the collection of data. Data from a candidate's credit record is analyzed using mathematical models for developing FICO credit scoring formula, which is not publicly disclosed. This method considers five important variables: the history of credit performance, the amount of debt owed at the moment, the length of credit history, the type of credit used, and most recent credit inquiries. FICO rating is more exact and considers 20 distinct elements in 5 categories. Payment history (35%), amount owed (30%), length of credit history (15%), type of credit (10%), and new applications (10%) are the five categories (https://www.myfico.com/credit-education/whats-in-your-credit-score). Similar methodology is used in the researched dataset, but some variables are substituted and some other extra predictors of the target variable are found. Most of the countries apply this type of credit scoring in credit scoring, required by law. There might be minor difference, but conceptual framework is similar, i.e., being able to determine the credibility of a credit applicant with the historical data, the only differences are minor differences, such as scoring scale and etc. (https://www.myfico.com/credit-education/fico-scores-vs-credit-scores).

### 1.2 Research Problem Statement

Logistic regression is another statistical technique widely utilized in credit scoring to evaluate the likelihood of default and assess the associated risks of making a decision (Gouvêa & Gonçalves, 2021, p.198). This approach is considered one of the most popular models for credit scoring due to its ability to effectively model the relationship between a set of independent variables and the probability that a given case belongs to a particular category of the dependent variable.

By modelling and analyzing the link between a categorical dependent variable and one or more independent factors, logistic regression enables us to gain insightful knowledge about binary or multinomial outcomes (Agresti, 2013, p.3). However, using conventional logistic regression can be cumbersome if a lot of categorical independent variables with numerous levels are being used. If this is the case the number of model's parameters becomes very large.

The Weight of Evidence (WoE) approach to credit scoring has drawn interest as a potential remedy for the shortcomings of conventional logistic regression models Recent research has indicated that the WoE approach may be able to enhance the stability and accuracy of credit scoring models, particularly when working with categorical predictor variables that have several levels. Additionally, it has been discovered that the method offers a more understandable and intuitive way of determining creditworthiness, allowing lenders to make more informed lending decisions. Since there hasn't been enough thorough research on the performance and applicability of WoE method in the context of credit scoring, the regulatory compliance of the WoE method in credit scoring models is still up for debate. In particular, it's important to look into how well the WoE technique handles categorical predictor variables with many levels and assess how it affects model correctness, interpretability, and stability. By presenting empirical data on the usefulness of the WoE approach in logistic regression for credit scoring and its implications for risk assessment and decision-making in the credit business, this study attempts to close the knowledge gap currently present in the field.

### 1.3 Research purpose

The goal of this study is to clarify whether financial institutions should be encouraged to use the WoE method technique when evaluating a potential borrower's creditworthiness, or if the use of less sophisticated predictive models in credit scoring determines the borrowing party's creditworthiness to an equivalent degree. Risky borrowers are those who, although appearing to have strong credit, are more likely to default on their loans. Therefore, the research question of whether the financial institutions should have an incentive towards integrating the WoE method in credit scoring is a valid and important one. By integrating the WoE method, financial institutions can potentially improve their credit evaluation process and reduce the risk of default by identifying high-risk borrowers who may be overlooked by traditional credit scoring models. However, it is important to note that the implementation of the WoE method may require additional resources and expertise from financial institutions. They may need to invest in data analytics software, train their employees in the WoE method, and develop new processes for integrating the WoE method into their existing credit evaluation process.

In conclusion, the purpose of this study is to investigate the viability and efficacy of incorporating the WoE method into the credit scoring process. The results of this study may help financial organizations decide whether to switch to the WoE technique or stick with their traditional credit scoring algorithms.

# 2. Logistic regression and Weight of Evidence

## 2.1 Logistic regression

Financial organizations can use logistic regression to basically forecast the likelihood that a borrower will fail on a loan based on a number of variables, including credit history, income, employment status, and other pertinent data points. A logistic regression model may use these characteristics to create a risk score that indicates the possibility of default. This score can then be used to guide lending decision, determine the right interest rates, and pinpoint high-risk borrowers (Thomas, Edelman & Crook, 2002, p.16).

One of the key advantages of logistic regression is its ability to accommodate binary outcomes, where the dependent variable takes on one of two possible values (Douglas, 2005, p.1545). This makes it particularly well-suited for credit scoring, where the goal is often to predict whether a borrower will default or not. Additionally, logistic regression can also handle categorical or ordinal variables, making it a versatile tool for analyzing a range of credit-related data.

Overall, logistic regression plays a vital role in credit scoring, enabling financial institutions to make informed lending decisions, mitigate risk, and ensure the sustainability of their business operations. As the use of big data and machine learning continues to grow in the financial sector, it is likely that logistic regression will remain a critical tool for assessing credit risk and making informed lending decisions. We next describe the formal model of logistic regression.

Logistic regression models the probability that the response (or target) variable Y belongs to a particular category. For example, in our study, the response variable Y is '*status*' of the loan recipient and the category of interest is the values Y=1 that corresponds to 'good' recipients of loan.

Let $X_1, \dots, X_p$ be the independent variables used in predicting the probability of the event Y=1. For convenience and shortness, let us denote the vector $X = (X_0, X_1, \dots, X_p)'$, where $X_0$ is a formal constant variable, $X_0 \equiv 1$. The conditional probability of Y=1 given the values of $X_1, \dots, X_p$ is denoted by $\pi(X) = P(Y = 1|X)$. In logistic regression, not the probability $\pi(x)$ itself but its logit transform $\ln\left(\frac{\pi(X)}{1-\pi(X)}\right)$ is modeled. The reason is that when trying to model $\pi(X)$ itself, the simple regression model can easily produce probabilities outside interval [0, 1], which is not acceptable. In contrast, the logit transform $\ln\left(\frac{\pi(X)}{1-\pi(X)}\right)$ allows for arbitrary (including negative) values on the real line. Therefore, in logistic regression, the logit transform of $\pi(X)$ is modeled as a linear combination of the independent variables $X_1, \dots, X_p$:

$$\ln\left(\frac{\pi(X)}{1-\pi(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p. \quad (1)$$

The regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ are the parameters of the model which are to be estimated from the data. By introducing a column vector $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$ and recalling that $X = (X_0, X_1, \ldots, X_p)'$, the equation (1) can be written in a more compact form:

$$\ln\left(\frac{\pi(X)}{1-\pi(X)}\right) = \beta' X. \quad (2)$$

By solving the equation (2) for $\pi(X)$, one easily obtains that

$$\pi(X) = \frac{e^{\beta' X}}{1+ e^{\beta' X}}. \quad (3)$$

This equation implies that the probability of the eve Y=1 changes when the value of the independent variables vary. By using this equation, researches can acquire insights into how various factors affect the probability $\pi(X)$.

### 2.2 Estimation in logistic regression

Maximum likelihood estimation (MLE) is used in logistic regression to estimate model parameters based on observed data. The goal is to determine the collection of parameter values that maximizes the model's likelihood of the observed data.

Let's go over the steps in MLE for logistic regression:

1. Establish the Logistic Regression Model:

a. The logistic regression model assumes a linear relationship between the independent variables (X) and the log-odds (logarithm of the odds) of the dependent variable (Y).

b. The log-odds of Y being in a certain category (Y=1, in our case) is represented by the equation (2) where $\pi(X)$ represents the probability of Y being in category 1 given the vector *X* of independent variables, and β is the vector of unknown coefficients to be estimated.

2. Determine the Likelihood Function:

a. The data consists i.e., N observations of vectors $X^{(1)}, \ldots, X^{(N)}$ and their respective responses $Y_1, \ldots, Y_N$ .

b. The likelihood function represents the probability of observing the given data given the parameter vector $\beta$.

c. The likelihood function is derived based on the assumption that the observations of the response variable are independent and follow a Bernoulli distribution.

d. For each observation i, the probability of Y=1 given $X^{(i)}$ is represented by equation (3) and is denoted by

$$\pi_i = \pi(X^{(i)}) = \mathrm{P}(\mathrm{Y}=1|X^{(i)}) = \frac{e^{\beta' X^{(i)}}}{1+e^{\beta' X^{(i)}}}. \quad (4)$$

e. The likelihood function is the product of these probabilities across all observations:

$$\mathrm{L}(\beta) = \prod_{i=1}^{N} {\pi_i}^{Y_i} (1-\pi_i)^{1-Y_i},$$

where $\mathrm{Y_i}$ is the observed category of Y for the i-th observation ($Y_i = 1$ or 0).

3. Maximize the Likelihood Function:

To estimate the parameter values that maximize the likelihood function, we take the natural logarithm of the likelihood function (lnL(β)) and aim to find the β that maximize lnL(β). Taking the logarithm helps simplify the calculations and also transforms the product into a sum terms. The expression

$$\mathrm{lnL}(\beta) = \sum_{i=1}^{N} [Y_i \ \ln(\pi_i) + (1 - Y_i) \ \ln(1 - \pi_i))]$$

represents the log-likelihood function.

4.Optimizations:

The likelihood equations, obtained by differentiating lnL(β) with respect to $\beta_0, \beta_1, \dots, \beta_p$ and setting the results equal to zero, are:

$$\sum_{i=1}^{N} (Y_i - \pi_i) = 0$$

and

$$\sum_{i=1}^{N} X_j^{(i)} (Y_i - \pi_i) = 0, \qquad \mathrm{j} = 1, \dots, \mathrm{p}\,.$$

Note that, because of (4), these equations are non-linear (with respect to beta coefficients) and require special methods for their solution.

a. Therefore, maximum of the log-likelihood function is commonly accomplished through the use of numerical optimization methods such as gradient descent or the Newton-Raphson algorithm. The optimization process iteratively adjusts the parameter values until the log-likelihood function is maximized.

5.Solution:

Once the optimization process has converged, the estimated parameter values (β) that maximize

the log-likelihood function is obtained. These estimates are the logistic regression model's coefficients.

### 2.3 Weight of Evidence

Logistic regression is the most commonly used method to predict the probability of default in credit scoring. Nonetheless, using the traditional predictive models can be cumbersome if a lot of categorical variables with numerous levels are being used. The Weight of Evidence (WoE) method is a statistical technique used in the areas of risk assessment (Abdou, 2009, p.11403). WoE transforms the values of a variable into discrete categories and assigns to each separate category a distinct WoE value (Good,1950, p.44). The WoE score will be high if the category has a high percentage of defaulters in comparison to non-defaulters, which indicates that the category does a good job of separating defaulters from non-defaulters. (Lin & Hsieh, 2014, p.1). The use of WoE gained popularity in credit risk assessment in the 1980s and 1990s, as credit card companies and other lenders sought more sophisticated methods for evaluating the creditworthiness of applicants. WoE allowed lenders to transform categorical variables, education level, into a single continuous variable that could be easily incorporated into credit scoring models.

In the early 2000s, WoE gained traction in the field of financial analytics as a way to analyze and optimize customer segmentation and targeting strategies. By transforming categorical variables into continuous variables with WoE, marketers could gain insights.

The use of WoE involves a transformation of data that requires binning, which is a process that transforms a continuous or a categorical variable into set groups or bins (Zeng, 2014, p.3229). According to Siddiqi (2006, p.80), a good binning strategy follows the following guidelines. Correct binning strategy is:

1) Each bin should cover at least 5 percent of the data,
2) Missing values are binned separately,
3) Each particular bin cannot consist only from good customers or bad customers (i.e., 0's and 1's).

The main purpose of binning is assuring the monotonic relationship with the target variable. Calculation of the WoE is done in the following way. Assume we have a dataset containing N observations and, let Y be the target variable, that is binary, i.e., Y=1 or Y=0.

Let $X_1 \ldots X_p$ be the set of independent variables. Let $B_1 \ldots B_k$ be bins for the variable $X_j$. The WoE for the variable $X_j$ in bin $i$ is then defined as

$$WOE_{ij} = \log\left(\frac{P(X_j \in B_i|Y=1)}{P(X_j \in B_i|Y=0)}\right)$$

where,

$$P(X_j \in B_i|Y=1) = \frac{N_{X_j \in B_i|Y=1}}{N_{X_j|Y=1}} = \frac{Total\ number\ of\ 'Good'applicants\ in\ bin\ B_i\ of variable\ X_j}{Total\ number\ of\ 'Good'applicants\ in\ variable\ X_j},$$

$$P(X_j \in B_i|Y=0) = \frac{N_{X_j \in B_i|Y=0}}{N_{X_j|Y=0}} = \frac{Total\ number\ of 'Bad'applicants\ in\ bin\ B_i\ \ of variable\ X_j}{Total\ number\ of\ 'Bad'applicants\ in\ variable\ X_j}.$$

=

Alternatively, the WoE can also be expressed in terms of odds ratio. Recall that odds are the probability of the event *(Y=1)* occurring divided by the probability of the event not occurring, i.e., the ratio $\frac{P(Y=1)}{P(Y=0)}$. Now, because of

$$WOE_{ij} = \log\left(\frac{P(X_j \in B_i|Y=1)}{P(X_j \in B_i|Y=0)}\right) = \log\left(\frac{P(X_j \in B_i, Y=1)}{P(X_j \in B_i, Y=0)} \Big/ \frac{P(Y=1)}{P(Y=0)}\right),$$

we see that $WOE_{ij}$ is the logarithm of the ratio of two odds, the first odds being calculated within the bin $B_i$, and the second being the general odds. Practical calculations of WoE-s are presented in subsection 4.2.

There are several reasons for using WoE. First, it should establish a monotonic relationship to the dependent variable (Baesens, Roesch, Scheule, 2016, p.116). However, non-monotonic relationship can occur, which can be kept as long as the relationship can be explained (Siddiqi, 2006, p.84). It also deals with missing values and outliers conveniently.

In practice, there is also possibility of the emerging drawbacks from using Weight of evidence method. First, there might occur a loss of information, which may happen after the application of binning procedure. Second, correlations among the explanatory variables are not taken into consideration, i.e., there may be a strong correlation between some independent variables, which highlights the importance of doing data exploration before applying the technique.

# 3 Data findings and logistic regression

## 3.1 Target population

The target population of the research are the individuals who wish to apply for a credit. Their credit worthiness is determined by the data obtained from them such as their age, expenditure, education level. Those very variables and many other variables help to identify the applicant's riskiness. By examining the relationship between the various factors, such as occupation, expenditure, and education level, and their impact on creditworthiness, it is possible to develop a predictive model that can be used to assess the creditworthiness of future applicants.

## 3.2 Description of Data

The data was obtained from a financial institution of Estonia. The sample consists of 3985 individuals. The given data includes 17 variables. The predicted variable is Status. If Status = 1, it indicates to a 'Good' customer; if Status = 0, it indicates to the 'Bad' customer.

The predictor variables are the individual's age, gender, family status, first language, region of origin, sum of credit, period of the credit, work experience, education level, number of individual's children, whether the individual possess any Estate object, income of an individual, expenditure of an individual. The data also includes the number of payment alerts applicant ever had, and how many of them are active and how many of them are closed. The predictor variables are given in the following table.

| Variable | Description | Type of Variable |
|---|---|---|
| $X_1$ | Sex | Nominal |
| $X_2$ | Age | Numerical |
| $X_3$ | Region | Nominal |
| $X_4$ | Language | Nominal |
| $X_5$ | Sum | Numerical |
| $X_6$ | Log(Period) | Numerical |
| $X_7$ | Outcome | Numerical |
| $X_8$ | Income | Numerical |
| $X_9$ | Children | Numerical |
| $X_{10}$ | Estate | Numerical |
| $X_{11}$ | Active Payment Alerts | Numerical |
| $X_{12}$ | Closed Payment Alerts | Numerical |
| $X_{13}$ | Total Payment Alerts | Numerical |
| $X_{14}$ | Family | Nominal |
| $X_{15}$ | Education | Nominal |
| $X_{16}$ | Work Experience | Nominal |

The values of the nominal variables are given in the following table:

| Variable | Values |
|---|---|
| Sex | Female, Male |
| Region | Harjumaa, Tartumaa, Ida-Virumaa, other |
| Family | Single, Married, Divorced, Cohabiting, Widowed |
| Education | No Education, Basic, Higher, Vocational, Secondary |
| Work Exp. | Unemployed, < 1 year, > 1 year, Intern |

Certain data manipulations have been applied, such that, for example, individuals in our sample are mainly from Harjumaa (around 55%), Tartumaa (around 9.5%), Ida-Virumaa (around 15%). The next highest populated region of Estonia is Parnumaa (around 3%). Hence, in the dataset, the Region variable will have 4 categories (Harjumaa, Tartumaa, Ida-Virumaa and all other regions of Estonia) instead of 15 original categories/.

### 3.3 Modeling strategy

In the following research we estimate four logistic regression models: model0 and 1 use original covariates and models 2 and 3 use WoE transformed covariates.

| Model equations | Variable selection methodology | WoE transformed variables |
|---|---|---|
| Model 0 | Full model (No selection) | NO |
| Model 1 | Backward selection | NO |
| Model 2 | No selection | YES |
| Model 3 | Backward selection | YES |

After having estimated all four models, one can decide about the effectiveness of the WoE transformation by comparing the quality of Models 2 and 0, and Models 3 and 1. Models 0 and 1 are regarded as the benchmark models.

### 3.4 Modeling logistic regression equation

The following analysis was done by using the RStudio software (version 4.2.2). First, a logistic regression model was estimated from 3985 data observations to obtain a credit scoring model. Out of 3985 sample points, 3983 observations were able to be a subject for examination, on account of 2 unreadable by software attributes in ‘Total Payment Alerts’ covariate. The dependent variable Y is ‘Status’ variable. ‘1’ indicates to a ‘Good’ customer, and ‘0’ indicating a ‘Bad’ counterpart.

| Status | |
|---|---|
| Y = 1<br>(‘Good customers’) | 2823 |
| Y = 0<br>(‘Bad customers’) | 1160 |
| Total | 3983 |

Another feature in credit-scoring modeling that has been studied is variable selection. The variables of WoE transformed logistic regression is also selected via Backward or Forward Selection. (Abdou & Pointon, 2011, p.66). We start with the estimation of the ‘full’ Model 0, and then proceed to the Model 1, where the variable selection is done using backward selection strategy.

### 3.4.1 Model 0

Model 0 is the logistic regression equation where all possible predictors are present. Only the term that corresponds to’ *total payment alerts* was excluded, because total payment alerts is sum of *active alerts* and *closed alerts*.

In case of other nominal variables (sex, language, family status work experience), one of the values of each variable was defined as the baseline value (i.e., reference category), and other values were included in the model using indicator functions. For example, the value ‘female’ of the sex variable is used as a baseline value, but ‘male’ is included via the indicator

$$I = \begin{cases} 1, & if\ sex = 'male' \\ 0, & if\ sex = 'female' \end{cases}$$

| **Model 0** | | | | | | | |
|---|---|---|---|---|---|---|---|
| Variables and Indications | Labels | Beta coef’s | S.E[a] | Wald[b] | D.F.[c] | P-val | O.R[d] |
| Intercept | C | 0.9782 | 0.2511 | 15.2 | 1 | 0.0001 | 2.65967 |
| Sex: Male | $X_1$ | -0.3685 | 0.0839 | 19.3 | 1 | 0.0000 | 0.69177 |
| Age | $X_2$ | 0.017 | 0.0039 | 19.3 | 1 | 0.0000 | 1.01715 |
| Region:Other | $X_3$ | -0.272 | 0.1042 | 6.8 | 1 | 0.0091 | 0.76185 |
| Region: Ida-Virumaa | $X_4$ | -0.3586 | 0.1129 | 10.1 | 1 | 0.0015 | 0.69865 |
| Region: Tartumaa | $X_5$ | -0.137 | 0.1329 | 1.1 | 1 | 0.3023 | 0.87197 |
| Language:Rus | $X_6$ | 0.1542 | 0.0899 | 2.9 | 1 | 0.0863 | 1.16672 |
| Sum | $X_7$ | -0.0004 | 0.0002 | 4.5 | 1 | 0.0329 | 0.9996 |
| Log(Period) | $X_8$ | -0.1533 | 0.044 | 12.1 | 1 | 0.0005 | 0.85787 |
| Outcome | $X_9$ | 0.0005 | 0.0002 | 4.8 | 1 | 0.0285 | 1.0005 |
| Income | $X_{10}$ | 0.0000 | 0.0001 | 0.0051 | 1 | 0.9433 | 1 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Children | $X_{11}$ | -0.093 | 0.0462 | 4.1 | 1 | 0.0441 | 0.91119 |
| Estate | $X_{12}$ | 0.5905 | 0.0627 | 88.8 | 1 | 0.0000 | 1.80489 |
| Active P.A. | $X_{13}$ | -0.2148 | 0.042 | 26.1 | 1 | 0.0000 | 0.8067 |
| Closed P.A. | $X_{14}$ | -0.0323 | 0.0161 | 4 | 1 | 0.0448 | 0.96822 |
| Family:Married | $X_{15}$ | 0.0588 | 0.1119 | 0.28 | 1 | 0.5991 | 1.06056 |
| Family: Divorced | $X_{16}$ | 0.0516 | 0.1535 | 0.11 | 1 | 0.7365 | 1.05295 |
| Family: Widowed | $X_{17}$ | -0.1162 | 0.2566 | 0.21 | 1 | 0.6506 | 0.8903 |
| Family: Cohabiting | $X_{18}$ | -0.0693 | 0.1027 | 0.46 | 1 | 0.4994 | 0.93305 |
| Education:Primary | $X_{19}$ | -0.1299 | 0.3992 | 0.11 | 1 | 0.7449 | 0.87818 |
| No Education | $X_{20}$ | -0.2772 | 0.3181 | 0.76 | 1 | 0.3834 | 0.7579 |
| Education: Higher | $X_{21}$ | 0.3598 | 0.1181 | 9.3 | 1 | 0.0023 | 1.43304 |
| Education: Vocational | $X_{22}$ | 0.1245 | 0.0903 | 1.9 | 1 | 0.1678 | 1.13258 |
| Education: Basic | $X_{23}$ | -0.3103 | 0.1216 | 6.5 | 1 | 0.0107 | 0.73323 |
| Work Exp.:Intern | $X_{24}$ | -0.1446 | 0.2265 | 0.41 | 1 | 0.523 | 0.86537 |
| Work Exp.: < 1 year | $X_{25}$ | -0.4109 | 0.0966 | 18.1 | 1 | 0.0000 | 0.66305 |
| Unemployed | $X_{26}$ | 0.1385 | 0.1897 | 0.53 | 1 | 0.4653 | 1.14855 |

**Clarification of title/labels assigned to the columns:**

a. S.E. – These are the standard errors for the coefficients. The standard errors aid in determining the statistical significance of the parameters.

b. Wald –z-score of Wald chi-square test. Using the Wald chi-square test, hypothesis that the beta coefficients is equal to 0 is examined.

c. D.F. – a measure of the Wald chi-square test's degree of freedom. There is a single df provided for each variable.

d. O.R. – odds ratio, exponentiated form of beta coefficients. Epyodds ratio is a statistical metric used to assess the degree of correlation between two variables. It measures the ratio of the likelih ood of an event happening in one group as opposed to another. There is no distinction in the grou ps' probability of event occurring when the value is 1. A value greater than 1 denotes that the eve nt is more likely to take place in the second group.

**Interpretation of Model 0**

If we look at Model 0, we can observe that 2 groups of variables and 1 numeric variable lack in statistical significance. One of the statistically insignificant variables is the *language* spoken by the applicant ($X_6$) (p-value = 0.0863). The variable itself can take 2 values, "*RUS*" and "*EST*". Reference category is the "EST" category. The other insignificant variable is the *family status*, consisting of 5 different categories: married, divorced, single, widowed, cohabiting ($X_{15}$, …, $X_{18}$). In our Model 0 with all covariates the intercept $\beta_0$ is the natural logarithm of odds ($\frac{P(Y=1)}{1-P(Y=1)}$) for

female applicant from Harjumaa county of Estonia, whose first language is Estonian, and who happens to be single, with secondary level of education and with more than 1 year of work experience (baseline values of respective variables). Using the odds ratio, we calculated in our Model 0, we see that the odds ratio for male applicant is 0.69177 (=exp(-0.3685)), meaning that if the applicant is a male the odds of dependent variable being equal to 1 decreases by 0.6917 (or, in the other words, the odds decrease approximately by 30%), given all other variables being unchanged (*ceteris paribus* principle). Let's also interpret the coefficient coming from a categorical explanatory variable like *family status.* The reference category for *family status* is '*single*' and we want to know how being married changes the odds of the applicant being creditworthy (i.e., probability of Y=1). From our model we can see that the odds ratio value for a married applicant is 1.06056. It means that, if the applicant is married, the odds of dependent variable being equal to 1, increases 1.0656 times (or, in other words, the odds increase approximately by 6%). Next, we interpret model parameters of a numeric variable like age. The beta coefficient of age variable is 0.017 and the odds ratio is 1.01715, which means that unit increase in age, (an keeping other variables unchanged) causes the log odds of applicant being equal to1 to increase for 0.17 (and hence the odds for the customer to be credit -worthy increase by 1.7%).

**Hosmer-Lemeshow table**

Hosmer-Lemeshow table is a model classification table which describes both expected model classifications and actual model classification. The Hosmer-Lemeshow splits the data in 10 groups (deciles, one per row) each representing the expected and observed frequency of both 1 and 0 values (Hosmer, Lemeshow & Sturdivant, p. 169). The expected frequency of data assigned to each decile should match the actual frequency outcome and each decile should contain data.

**Partition of Hosmer-Lemeshow**

| Group | Total | Y = 1 | | Y = 0 | |
|---|---|---|---|---|---|
| | | **Observed** | **Expected** | **Observed** | **Expected** |
| 1 | 398 | 248 | 277.75 | 150 | 120.25 |
| 2 | 398 | 255 | 281.47 | 143 | 116.53 |
| 3 | 398 | 262 | 279.66 | 136 | 118.34 |
| 4 | 398 | 267 | 278.69 | 131 | 119.31 |
| 5 | 398 | 275 | 279.54 | 123 | 118.46 |
| 6 | 398 | 295 | 286.18 | 103 | 111.82 |
| 7 | 398 | 305 | 287.42 | 93 | 110.58 |
| 8 | 398 | 313 | 282.85 | 85 | 115.15 |

| | | | | | |
|---|---|---|---|---|---|
| 9 | 398 | 296 | 283.85 | 102 | 114.15 |
| 10 | 398 | 307 | 285.78 | 91 | 112.22 |

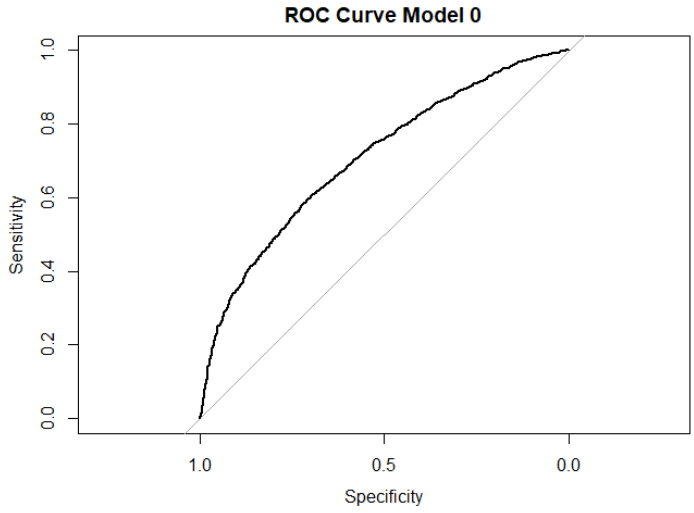
**Hosmer-Lemeshow Test (Model 0)**

| Chi-square | Degrees of freedom | p-value |
|---|---|---|
| 9.0009 | 8 | 0.3415 |

A logistic regression model's goodness of fit is evaluated using the H-L statistic. Based on the predicted probabilities from the model, it compares the observed and expected frequencies of outcomes over a collection of groups or bins. The Hosmer-Lemeshow test generates a chi-square statistic and a p-value that show how well or how poorly the model fits the data. A high p-value (higher than 0.05) indicates that the model fits the data well, whereas a low p-value (less than 0.05) indicates that the model does not. If the p-value from the Hosmer-Lemeshow test is 0.3415, as in our case, it suggests that the model fits the data well. It indicates that the model is good fit for the data, and the predicted probabilities from the model are consistent with the observed outcomes.

**ROC Curve (Model 0).**

A Receiver Operating Characteristic (ROC) curve is a graphical representation of the performance of a binary classification model, which is a model that classifies examples into one of two possible categories. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at different classification thresholds**.** The TRP is calculated with the ratio of true positive on total number of positive examples, FRP is calculated with the ratio of false positive on general number of positives.



**AUC (Area Under the ROC Curve)**

We calculate the AUC (Area Under the ROC Curve) as a metric to evaluate the overall performance of a binary classifier. The AUC represents the probability that a randomly selected positive example is ranked higher than a randomly selected negative example by the classifier. It is calculated by finding the area under the ROC curve using numerical integration or other methods. The AUC ranges from 0 to 1, with a higher AUC indicating better classifier performance. This variable should be as high as possible with some restrictions. Typical values indicate the following:

◦ 0.5 – No distinguish ability (the model has no meaning).

◦ 0.51 – 0.7 – Low distinguish ability (not a very good model yet the model can be used).

◦ 0.71 – 0.9 – Very good distinguish ability.

◦ 0.91 – 1 – Excellent distinguish ability.

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

**Area Under the Curve (Model 0)**

| Area | Standard error | P-value | Confidence Interval (95%) | |
|---|---|---|---|---|
| | | | Lower Bound | Upper bound |
| 0.7058 | 0.0872 | 0.00 | 0.6887 | 0.7229 |

We can conclude that, full model has a good distinguish ability. At the same time standard error of AUC is rather small (0.0872), due to considerably large sample size.

### 3.4.2 Model 1

In the upcoming model, the covariates which performed lack of statistical significance in Model 0, after applying step-wise backward selection were removed from the model equation gradually. The step-wise backward selection procedure was applied in order to see whether the beta coefficients will behave differently, if higher p-value coefficients are tossed gradually from the equation.

The following model will not include the covariates such as 'Language' (indicating to an individual whose first language is Russian) and 'Family' (indicating to the family status of an applicant), and numerical variable 'Income'.

| **Model 1** | | | | | | | |
|---|---|---|---|---|---|---|---|
| Variables and Indications | Labels | Beta coef's | S.E[a] | Wald[b] | D.F.[c] | P-val | O.R[d] |
| Intercept | C | 1.025 | 0.23752 | 18.6 | 1 | 0.0001 | 2.7871 |
| Sex: Male | $X_1$ | -0.3549 | 0.081 | 19.2 | 1 | 0.0000 | 0.7013 |
| Age | $X_2$ | 0.0178 | 0.0034 | 27 | 1 | 0.0000 | 1.018 |
| Region:Other | $X_3$ | -0.3416 | 0.0978 | 12.2 | 1 | 0.0005 | 0.7106 |
| Region: Ida-Virumaa | $X_4$ | -0.2966 | 0.1081 | 7.5 | 1 | 0.0061 | 0.7433 |
| Region: Tartumaa | $X_5$ | -0.1959 | 0.1287 | 2.3 | 1 | 0.1280 | 0.8221 |
| Sum | $X_6$ | -0.0004 | 0.0002 | 4.4 | 1 | 0.0362 | 0.9996 |
| Log(Period) | $X_7$ | -0.1549 | 0.0439 | 12.5 | 1 | 0.0004 | 0.8565 |
| Outcome | $X_8$ | 0.0004 | 0.0002 | 5.7 | 1 | 0.0167 | 1.0004 |
| Children | $X_9$ | -0.0892 | 0.0428 | 4.3 | 1 | 0.0372 | 0.9147 |
| Estate | $X_{10}$ | 0.5911 | 0.0626 | 89.1 | 1 | 0.0000 | 1.806 |
| Active P.A. | $X_{11}$ | -0.2131 | 0.042 | 25.7 | 1 | 0.0000 | 0.8081 |
| Closed P.A. | $X_{12}$ | -0.0348 | 0.016 | 4.7 | 1 | 0.0295 | 0.9658 |
| Education:Primary | $X_{13}$ | -0.075 | 0.3976 | 0.036 | 1 | 0.8504 | 0.9278 |
| No Education | $X_{14}$ | -0.2869 | 0.3168 | 0.82 | 1 | 0.3651 | 0.7506 |
| Education: Higher | $X_{15}$ | 0.3565 | 0.1177 | 9.2 | 1 | 0.0025 | 1.4283 |
| Education: Vocational | $X_{16}$ | 0.1422 | 0.0899 | 2.5 | 1 | 0.1136 | 1.1528 |
| Education: Basic | $X_{17}$ | -0.3217 | 0.1213 | 7.0 | 1 | 0.0080 | 0.7249 |
| Work Exp.:Intern | $X_{18}$ | -0.1603 | 0.226 | 0.5 | 1 | 0.4782 | 0.8519 |
| Work Exp.: < 1 year | $X_{19}$ | -0.4115 | 0.0965 | 18.2 | 1 | 0.0000 | 0.6627 |
| Unemployed | $X_{20}$ | 0.1411 | 0.1888 | 0.56 | 1 | 0.455 | 1.1515 |

**Interpretation of Model 1**

In the Model 1 where backward selection was used to exclude some insignificant covariates from the full model, the intercept coefficient $\beta_0$ is the natural logarithm of odds $(\frac{P(Y=1)}{1-P(Y=1)})$ for female a pplicants from Harjumaa county of Estonia, with secondary level of education and with work experience for more than 1 year. We see that the odds ratio for male applicant is 0.7013, hence the decrease is approximately 30% again, given all other variables being unchanged. Let's also interpret the coefficient coming from a categorical explanatory variable like *education.* The reference category for *education* is secondary level and we want to know how having higher changes the odds of the applicant being creditworthy (i.e., probability of Y=1). From our model we can see that the odds ratio value for an applicant with higher education is 1.4283. It means that, if the applicant has higher education, the odds of dependent variable being equal to 1 increase by 1.4283 (or, in other words, the odds increase approximately by 43%). Next, we

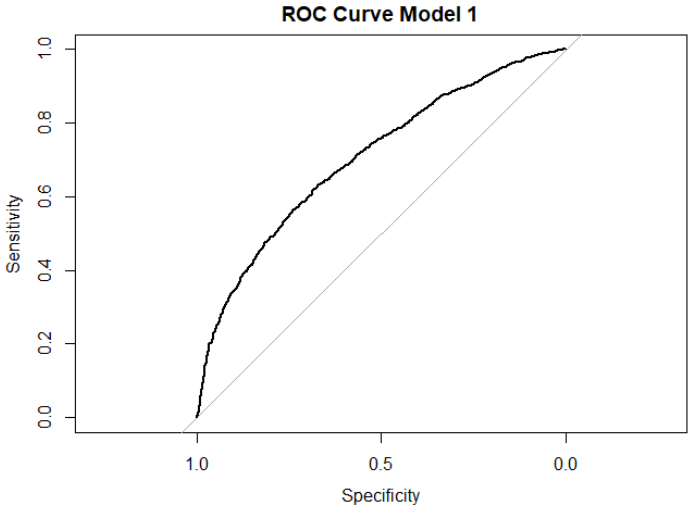
interpret model parameters of a numeric variable like age. The beta coefficient of age variable is 0.0178 and the odds ratio is 1.018, which means that unit increase in age, (an keeping other variables unchanged) causes the log odds of applicant being equal to1 to increase for 1.018 times (and hence the odds for the customer to be credit -worthy increase by 1.8%).

**Hosmer-Lemeshow test (Model 1)**

| Chi-square | Degrees of freedom | p-value |
|---|---|---|
| 8.7473 | 8 | 0.3641 |

H-L test for Model 1 showed slightly better result than H-L test for Model 0. It suggests that the observed data fits the expected model well. Therefore, we fail to reject null hypothesis that there is no significant difference between the observed and expected values. In other words, the H-L test confirms the goodness of fit of the logistic regression model.

**ROC Curve**




**Area Under the Curve (Model 1)**

| Area | Standard error | P-value | Confidence Interval (95%) | |
|---|---|---|---|---|
| | | | Lower Bound | Upper bound |
| 0.7046 | 0.088 | 0.00 | 0.6874 | 0.7217 |

An AUC of 0.7046 indicate that the logistic regression model has a moderately good ability to distinguish between the positive and negative outcomes. However, the model's predictive ability

is only moderately good. It is important to note that an AUC of 0.5 indicates a model with no predictive ability, while an AUC of 1.0 indicates to a perfect model. Therefore, the model may benefit from further WoE improvements to increase its predictive performance. In addition to improving the model's performance, it is also important to interpret the coefficients of the logistic regression model to gain insights into the relationship between the predictors and the outcome variable.

# 4 Logistic regression with WoE transformed variables

## 4.1 Variable binning strategy

In order to Apply WoE method to the logistic regression model, the variables should be binned, to make categorical variables continuous, and see how the relationship between the response variable changes. As it was mentioned earlier, correct binning strategy would be considered if every bin covers at least 5% of data (for our sample space it makes around 200 observations). Definitely, there is certain freedom in binning. I the ideal case, the bins are chosen which produce best possible quality indicators of the resulting logistic regression model. However, in this resulting work we did not pay much attention to the binning optimality problem.

## 4.2 Application of WoE Transformation

### WoE Transformation of 'Age' Covariate

Before applying WoE to numerical variables, we need to categorize them as well. "Age" variable is such an example. We divide *age* into following 9 categories.

| Age group | Bad | Good | Total | Bad% | Good% | WoE |
|---|---|---|---|---|---|---|
| 18-25 | 196 | 236 | 432 | 0.168966 | 0.083599 | -0.7037 |
| 25-30 | 218 | 351 | 569 | 0.187931 | 0.124336 | -0.4131 |
| 30-35 | 159 | 384 | 543 | 0.137069 | 0.136026 | -0.0076 |
| 35-40 | 167 | 401 | 568 | 0.143966 | 0.142047 | -0.0134 |
| 40-45 | 147 | 350 | 497 | 0.126724 | 0.123982 | -0.0219 |
| 45-50 | 108 | 310 | 418 | 0.093103 | 0.109812 | 0.16506 |
| 50-55 | 71 | 297 | 368 | 0.061207 | 0.105207 | 0.54167 |
| 55-60 | 56 | 260 | 316 | 0.048276 | 0.092101 | 0.6460 |
| 60-70 | 38 | 234 | 272 | 0.032759 | 0.082891 | 0.9284 |
| Total | 1160 | 2823 | 3983 | 1 | 1 | |

The last column (WoE) of the table represents the WoE transformed version of the variable *age*, being a discrete numerical variable with 9 values. It is seen that the relationship between the initial variable *age* and its WoE transformed version is more or less monotonic (with some disturbances in the middle classes).

**WoE transformation of 'Sum' covariate**

'Sum' variable is a numeric variable, which can be categorized according to correct binning to correct binning strategy that we apply. *Sum* is categorized into the following 7 groups.

| Sum group | Bad | Good | Total | Bad % | Good% | WoE |
|---|---|---|---|---|---|---|
| <=100 | 156 | 488 | 644 | 0.134483 | 0.172866 | 0.25107 |
| 100-200 | 284 | 620 | 904 | 0.244828 | 0.219625 | -0.10863 |
| 200-300 | 346 | 778 | 1124 | 0.298276 | 0.275593 | -0.07909 |
| 300-400 | 113 | 256 | 369 | 0.097414 | 0.090684 | -0.07159 |
| 400-500 | 142 | 331 | 473 | 0.122414 | 0.117251 | -0.04308 |
| 500 -800 | 74 | 190 | 264 | 0.063793 | 0.067304 | 0.05357 |
| 800-2000 | 45 | 160 | 205 | 0.038793 | 0.056677 | 0.37913 |
| Total | 1160 | 2823 | 3983 | 1 | 1 | |

Interestingly, now the relationship between the original variable 'Sum' and its WoE transformed version is U-shaped rather than monotonic.

**WoE transformation of 'Children' covariate**

The *children* variable is divided into 3 categories.

| Number of children | Bad | Good | Total | Bad% | Good% | WoE |
|---|---|---|---|---|---|---|
| 0 | 691 | 1774 | 2465 | 0.59569 | 0.628409 | 0.05347 |
| 1 | 266 | 630 | 896 | 0.22931 | 0.223167 | -0.02716 |
| Over 1 | 203 | 419 | 622 | 0.175 | 0.148424 | -0.16472 |
| Total | 1160 | 2823 | 3983 | 1 | 1 | |

**WoE transformation of 'Estate' covariate**

The *estate* variable is divided into 3 categories.

| Number of Estates | Bad | Good | Total | Bad% | Good% | WoE |
|---|---|---|---|---|---|---|
| 0 | 838 | 1344 | 2182 | 0.722414 | 0.476089 | -0.41699 |
| 1 | 268 | 1105 | 1373 | 0.231034 | 0.391428 | 0.52723 |
| Over 1 | 54 | 374 | 428 | 0.046552 | 0.132483 | 1.04589 |
| Total | 1160 | 2823 | 3983 | 1 | 1 | |

**WoE transformation of 'Active Payment Alerts' covariate**

The binning procedure for this variable was done in such a way that each bin covers at least 5% of the data. However, due to the distribution of the data, the binning procedure had to be divided into two separate bins. The first bin corresponds to individuals who have *active payment alerts*, while the second bin includes individuals who do not have *active payment alerts*.

| Active alerts | Bad | Good | Total | Bad% | Good% | WoE |
|---|---|---|---|---|---|---|
| No | 972 | 2546 | 3518 | 0.837931 | 0.901877 | 0.073542 |
| Yes | 188 | 277 | 465 | 0.162068 | 0.098122 | -0.501804 |
| Total | 1160 | 2823 | 3983 | 1 | 1 | |

**WoE transformation of 'Closed Payment Alerts' covariate**

Since, every bin should cover at least 5% of the data, binning procedure had to be divided into 5 bins.

| Closed alerts | Bad | Good | Total | Bad% | Good% | WoE |
|---|---|---|---|---|---|---|
| 0 | 588 | 1846 | 2434 | 0.506897 | 0.653914 | 0.254669 |
| 1 | 243 | 396 | 639 | 0.209483 | 0.140276 | -0.40102 |
| 2 | 136 | 216 | 352 | 0.117241 | 0.076514 | -0.42675 |
| 3 | 76 | 128 | 204 | 0.065517 | 0.045342 | -0.36808 |
| Over 3 | 117 | 237 | 354 | 0.100862 | 0.083953 | -0.18349 |
| Total | 1160 | 2823 | 3983 | 1 | 1 | |

**WoE transformation of 'Outcome' covariate**

In our analysis, we employ a binning strategy to categorize the *outcome* variable, which is a discrete numeric variable, into six groups based on their values.

| Outcome category | Bad | Good | Total | Bad% | Good% | WoE |
|---|---|---|---|---|---|---|
| <100 | 141 | 366 | 507 | 0.121552 | 0.129649 | 0.06449 |
| 100-200 | 358 | 703 | 1061 | 0.308621 | 0.249026 | -0.21455 |
| 200-300 | 265 | 602 | 867 | 0.228448 | 0.213248 | -0.06445 |
| 300-400 | 162 | 451 | 613 | 0.139655 | 0.159759 | -0.13449 |
| 400-500 | 105 | 302 | 407 | 0.090517 | 0.106978 | 0.16708 |
| >500 | 129 | 399 | 528 | 0.111207 | 0.141339 | 0.23977 |
| Total | 1160 | 2823 | 3983 | 1 | 1 | |

**WoE transformation of 'Region' covariate**

For applying WoE transformation applicants have been grouped into 4 categories

| Region | Bad | Good | Total | Bad% | Good% | WoE |
|---|---|---|---|---|---|---|
| Harjumaa | 573 | 1628 | 2201 | 0.493966 | 0.576691 | 0.154842 |
| Tartumaa | 118 | 264 | 382 | 0.101724 | 0.093518 | -0.08412 |
| I-V | 186 | 416 | 602 | 0.1600345 | 0.147361 | -0.08444 |
| other | 283 | 515 | 798 | 0.243966 | 0.18243 | -0.29066 |
| Total | 1160 | 2823 | 3983 | 1 | 1 | |

**WoE transformation of 'Family' covariate**

*Family* variable consists of 5 groups.

| Family | Bad | Good | Total | Bad% | Good% | WoE |
|---|---|---|---|---|---|---|
| Single | 416 | 805 | 1221 | 0.358621 | 0.285158 | -0.22922 |
| Married | 284 | 884 | 1168 | 0.244828 | 0.313142 | 0.246103 |
| Divorced | 90 | 329 | 419 | 0.077586 | 0.116543 | 0.406868 |
| Widowed | 26 | 101 | 127 | 0.022414 | 0.035778 | 0.467644 |
| Cohabiting | 344 | 704 | 1048 | 0.296552 | 0.24938 | -0.17324 |
| Total | 1160 | 2823 | 3983 | 1 | 1 | |

**WoE transformation of 'Education' covariate**

*Education* variable consists of 6 categories.

| Education | Bad | Good | Total | Bad% | Good% | WoE |
|---|---|---|---|---|---|---|
| Keskharidus | 467 | 1081 | 1548 | 0.402586 | 0.382926 | -0.05007 |
| Alglharidus | 11 | 20 | 31 | 0.009483 | 0.007085 | -0.29154 |
| No Educ | 17 | 32 | 49 | 0.014655 | 0.011335 | -0.25686 |
| Kõrgharidus | 133 | 593 | 726 | 0.114655 | 0.21006 | 0.605465 |
| Kutseharidus | 331 | 863 | 1194 | 0.285345 | 0.305703 | 0.068916 |
| Põhiharidus | 201 | 234 | 435 | 0.173276 | 0.082891 | -0.73736 |
| Total | 1160 | 2823 | 3983 | 1 | 1 | |

**WoE transformation of 'Work Experience' covariate**

*Work Experience* variable consists of 4 categories.

| W.E. | Bad | Good | Total | Bad% | Good% | WoE |
|---|---|---|---|---|---|---|
| >1 year | 801 | 2248 | 3049 | 0.690517 | 0.796316 | 0.142555 |
| <1 year | 275 | 381 | 656 | 0.237069 | 0.134963 | -0.56335 |
| Intern | 37 | 64 | 101 | 0.031897 | 0.022671 | -0.34141 |
| Unemployed | 47 | 130 | 177 | 0.040517 | 0.04605 | 0.128007 |
| Total | 1160 | 2823 | 3983 | 1 | 1 | |

**WoE transformation of 'Income' covariate**

*Income* is a numeric variable, which will be binned into 6 groups.

| Income category | Bad | Good | Total | Bad% | Good% | WoE |
|---|---|---|---|---|---|---|
| <400 | 159 | 389 | 548 | 0.137069 | 0.137797 | 0.005295 |
| 400-600 | 344 | 735 | 1079 | 0.296552 | 0.260361 | -0.13015 |
| 600-800 | 309 | 687 | 996 | 0.266379 | 0.243358 | -0.09039 |
| 800-1000 | 168 | 476 | 644 | 0.144828 | 0.168615 | 0.152074 |
| 1000-1200 | 72 | 206 | 278 | 0.062069 | 0.072972 | 0.16183 |
| >1200 | 108 | 330 | 438 | 0.093103 | 0.116897 | 0.227581 |
| Total | 1160 | 2823 | 3983 | 1 | 1 | |

**WoE transformation of 'log(Period)' covariate**

Period is a numeric variable. Log(*Period*) is log transformed version of the Period variable, which will be binned into 7 groups.

| Log(Period) | Bad | Good | Total | Bad% | Good% | WoE |
|---|---|---|---|---|---|---|
| <3.4 | 24 | 260 | 284 | 0.02069 | 0.092101 | 1.493248 |
| 3.4-3.5 | 369 | 1096 | 1465 | 0.318103 | 0.388239 | 0.199246 |
| 3.5-4.1 | 208 | 323 | 531 | 0.17931 | 0.114417 | -0.44927 |
| 4.1-4.5 | 176 | 267 | 443 | 0.151724 | 0.09458 | -0.47262 |
| 4.5-5.1 | 93 | 173 | 266 | 0.080172 | 0.061282 | -0.26869 |
| 5.1-5.8 | 169 | 340 | 509 | 0.14569 | 0.120439 | -0.19033 |
| >5.8 | 121 | 364 | 485 | 0.10431 | 0.128941 | 0.211983 |
| Total | 1160 | 2823 | 3983 | 1 | 1 | |

Next we will use the WoE transformed variables defined above as predictor variables in logistic regression models.

## 4.3 Modeling Logistic Regression Equation with WoE Transformed variables

### 4.3.1 Model 2

We now perform logistic regression modeling based on WoE transformed form are included. Differently from Model 0, In Model 2 each nominal variable is now represented by just one single WoE variable, which makes the total number of parameters significantly smaller than Model 0 (16 versus 27 parameters).

| **Model 2** | | | | | | | |
|---|---|---|---|---|---|---|---|
| Variables and Indications | Labels | Beta coef's | S.E[a] | Wald[b] | D.F.[c] | P-val | O.R[d] |
| Intercept | C | 1.0226 | 0.0673 | 231.2 | 1 | 0.0000 | 2.78041 |
| Sex: Male | $X_1$ | -0.3292 | 0.0816 | 17.3 | 1 | 0.0000 | 0.7195 |
| WoE: Age | $X_2$ | 0. 4886 | 0.1025 | 22.7 | 1 | 0.0000 | 1.63003 |

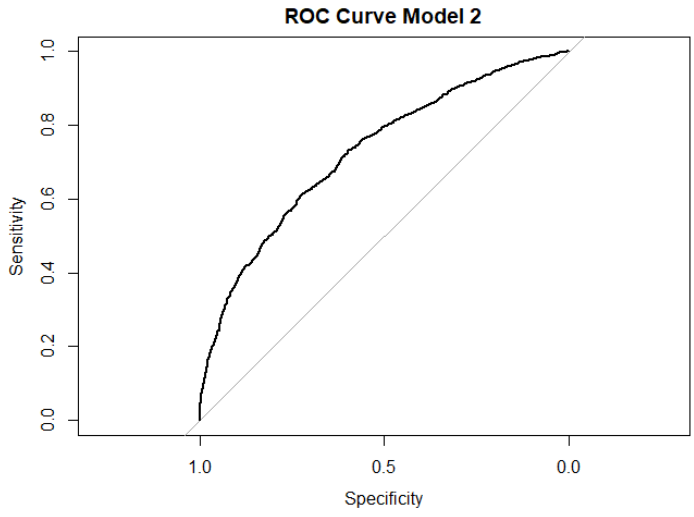
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| WoE: Region | $X_3$ | 0.5785 | 0.2211 | 6.8 | 1 | 0.0089 | 1.78336 |
| Language:Rus | $X_4$ | 0.0833 | 0.0796 | 1.1 | 1 | 0.2952 | 1.73568 |
| WoE: Sum | $X_5$ | 0.53828 | 0.264 | 4.2 | 1 | 0.0414 | 2.25376 |
| WoE: logPeriod | $X_6$ | 0.81175 | 0.0951 | 72.9 | 1 | 0.0000 | 1.16672 |
| WoE: Income | $X_7$ | 0.2061 | 0.329 | 0.39 | 1 | 0.5310 | 1.22888 |
| WoE: Outcome | $X_8$ | 0.848 | 0.2571 | 10.9 | 1 | 0.0010 | 2.33497 |
| WoE: Children | $X_9$ | 1.3426 | 0.5116 | 6.9 | 1 | 0.0087 | 3.82899 |
| WoE: Estate | $X_{10}$ | 0.7282 | 0.0782 | 86.6 | 1 | 0.0000 | 2.07135 |
| WoE: Active P.A. | $X_{11}$ | 0.4745 | 0.194 | 6.0 | 1 | 0.0145 | 1.60721 |
| WoE: Closed P.A. | $X_{12}$ | 0.6312 | 0.1241 | 25.9 | 1 | 0.0000 | 1.87987 |
| WoE: Family | $X_{13}$ | -0.0205 | 0.1745 | 0.014 | 1 | 0.9062 | 0.97971 |
| WoE: Education | $X_{14}$ | 0.3955 | 0.1114 | 12.6 | 1 | 0.0004 | 1.48513 |
| WoE: Work Experience | $X_{15}$ | 0.4461 | 0.1363 | 10.7 | 1 | 0.0011 | 1.56221 |

**Hosmer-Lemeshow test Model 2**

| Chi-square | Degrees of freedom | p-value |
|---|---|---|
| 12.393 | 8 | 0.1345 |

The p-value of 0. 1345 in the Hosmer-Lemeshow test indicates that there is insufficient evidence to reject the null hypothesis. The null hypothesis assumes that the model fits the data well, and the alternative hypothesis assumes that the model is a poor fit. But it is important to note that the Hosmer-Lemeshow test has limitations and should not be used as the sole method for evaluating the goodness of fit of a logistic regression model. Other methods, such as the ROC curve and calibration plot, can provide additional insight into the model's performance and help to identify any potential issues. It is recommended to use a combination of diagnostic tests to assess the validity of the model.

**ROC Curve Model 2**



**Area Under the Curve (Model 2)**

| Area | Standard error | P-value | Confidence Interval (95%) | |
|---|---|---|---|---|
| | | | Lower Bound | Upper bound |
| 0.7258 | 0.0085 | 0.00 | 0.7091 | 0.7425 |

After applying to 'all-covariate' model WoE variable/term transformation we are able to see that confidence interval changed quite noticeably. AUC also increased, comparing to model with untransformed variables, but the extent of significance will be known after "DeLong" test is applied, for knowing if the increase is statistically significant. It compares diagnostic evaluation accuracy by analyzing their ROC curves, assessing differences in performance (Delong E., Delong D., Clarke-Pearson, 1988, p.838). Nonetheless, 0.7258 AUC indicates to a decent model performance.

Now, to know the statistical significance of AUC value's change, we proceed to perform the DeLong test for two correlated ROC curves**.**

**DeLong test for the ROC curves of Model 0 and Model 2**

| DeLong test for two correlated ROC curves (all covariates) | | | |
|---|---|---|---|
| AUC of MODEL 0 | | AUC of MODEL 2 | |
| 0.7058 | | 0.7258 | |
| Z-score | P-values | Confidence interval of the difference | |
| -3.8617 | 0.0001126 | -0.030097188 | -0.009831599 |

As it was earlier mentioned, DeLong test is a statistical method used to compare the AUC of two correlated ROC curves. The test is used to determine if there is a statistically significant difference between two curves. In this case, a p-value of 0.0001126 is very low value, which means that the probability of observing the difference between the two curves due to chance is very small. Typically, a p-value of less than 0.05 is considered statistically significant, which means that the observed difference is unlikely to have occurred by chance performance (Delong E., Delong D., Clarke-Pearson, 1988, p.840). Therefore, in this case, the p-value of 0.0001126 suggests strong evidence against the null hypothesis that there is no difference between the two ROC curves, and supports the alternative hypothesis that the two curves are significantly different. In other words, the result of the DeLong test indicates that there is a statistically significant difference between the two ROC curves.

### 4.3.2 Model 3

The following model is the model with transformed variables, and with backward selection procedure applied

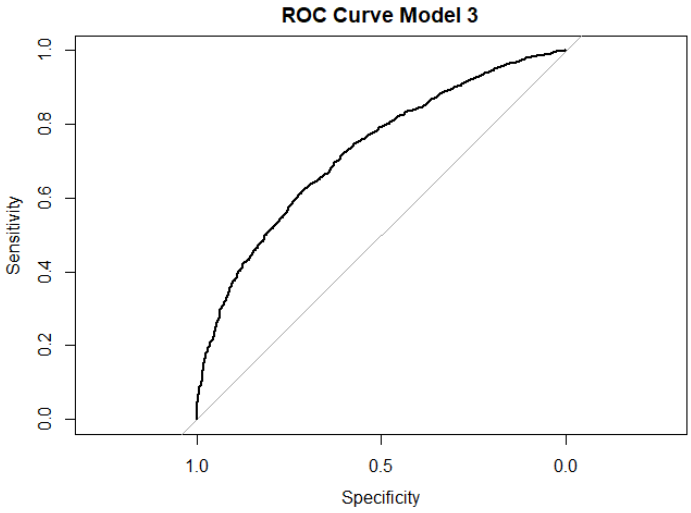
| Model 3 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Variables and Indications | Labels | Beta coef's | S.E[a] | Wald[b] | D.F.[c] | P-val | O.R[d] |
| Intercept | C | 1.0549 | 0.0567 | 346.7 | 1 | 0.0000 | 2.87169 |
| Sex: Male | $X_1$ | -0.3297 | 0.0797 | 17.1 | 1 | 0.0000 | 0.71914 |
| WoE: Age | $X_2$ | 0. 4838 | 0.0930 | 27.0 | 1 | 0.0000 | 1.61983 |
| WoE: Region | $X_3$ | 0.6466 | 0.2129 | 9.2 | 1 | 0.0024 | 1.90904 |
| WoE: Sum | $X_4$ | 0.5514 | 0.2634 | 4.4 | 1 | 0.3630 | 1.73568 |
| WoE: logPeriod | $X_5$ | 0.8126 | 0.0950 | 73.2 | 1 | 0.0000 | 2.25376 |
| WoE: Outcome | $X_6$ | 0.8872 | 0.2373 | 14.0 | 1 | 0.0002 | 2.42832 |
| WoE: Children | $X_7$ | 1.3436 | 0.4984 | 7.3 | 1 | 0.0070 | 3.83282 |
| WoE: Estate | $X_8$ | 0.7301 | 0.0779 | 87.9 | 1 | 0.0000 | 2.07529 |
| WoE: Active P.A. | $X_9$ | 0.4688 | 0.1939 | 5.8 | 1 | 0.0156 | 1.59808 |
| WoE: Closed P.A. | $X_{10}$ | 0.6345 | 0.1239 | 26.2 | 1 | 0.0000 | 1.88608 |
| WoE: Education | $X_{11}$ | 0.4052 | 0.1106 | 13.4 | 1 | 0.0002 | 1.4996 |
| WoE: Work Experience | $X_{12}$ | 0.4526 | 0.1361 | 11.1 | 1 | 0.0009 | 1.5724 |

**Hosmer-Lemeshow test Model 3**

| Chi-square | Degrees of freedom | p-value |
|---|---|---|
| 9.4778 | 8 | 0.3036 |

When the Hosmer-Lemeshow p-value is 0.3036, it suggests that the logistic regression model fits the data, and there is no significant difference between the observed and expected outcomes. This indicates that the model is a fit for the data and is able to accurately predict the likelihood of the outcome variable based on the predictor variables included in the model. Furthermore, the H-L test for the second WoE transformed variable model gave noticeably better fit result than in first WoE transformed variable model, suggesting that the second model is a more accurate representation of the data and may be a better predictor of the outcome variable than the first model.

**ROC Curve Model 3**




**Area Under the Curve (Model 3)**

| Area | Standard error | P-value | Confidence Interval (95%) | |
|---|---|---|---|---|
| | | | Lower Bound | Upper bound |
| 0.7253 | 0.009 | 0.00 | 0.7086 | 0.742 |

For our last model WoE transformed variable model, we can observe the same behavior, i.e. we see that AUC is higher than the AUC of corresponding model with no variable transformation. To know the statistical significance of AUC value's change, we will proceed to perform the 'DeLong' test for two correlated ROC curves**.**

**DeLong test for the ROC curves of Model 1 and Model 3**

| DeLong test for two correlated ROC curves (selected covariates) | | | |
|---|---|---|---|
| AUC of MODEL 1 | | AUC of MODEL 3 | |
| 0.7046 | | 0.7253 | |
| Z-score | P-values | Confidence interval of the difference | |
| -4.0409 | 0.00005324 | -0.03076730 | -0.01066943 |

A p-value of 0. 00005324 is an extremely low value, which means that the probability of observing the difference between the two curves due to chance is very small.

# Conclusion

In the context of the master's thesis, the Weight of Evidence (WoE) methodology in conjunction with logistic regression is a useful way for credit scoring. This methodology offers a solid foundation for determining the proper weights to give each credit risk factor and evaluating the predictive capability of different credit risk variables.

By employing logistic regression, the relationship between the credit risk factors and the likelihood of default accurately was modeled. The WoE transformation allowed to convert the numeric variables into discrete categories, making them more interpretable and facilitating the identification of non-linear relationships between the variables and the target variable.

One of the primary benefits of the WoE methodology is its ability to efficiently handle categorical data. It was accomplishable to capture the relative relevance and predictive potential of multiple categories inside a variable by computing the WoE values for each category. This method not only improves the model's interpretability but also allows for simple comparisons between different categories and variables.

The findings confirmed the usefulness of the WoE methodology in credit scoring. The logistic regression model with WoE predicted credit default risk accurately, allowing for effective risk assessment and decision-making. The performance of the model was validated using relevant evaluation metrics such as accuracy, precision, recall, and the area under the Receiver Operating Characteristic curve (AUC-ROC), which demonstrated its higher predictive potential.

In conclusion, the WoE methodology, when paired with logistic regression, provides a reliable and interpretable approach to credit scoring. Its capacity to handle categorical variables and provide insights into the relative relevance of different aspects makes it an important tool in the financial industry for credit risk assessment.

# References


1) Abdou, H., 2009, "Genetic programming for credit scoring: The case of Egyptian public sector banks", *Expert Systems with Applications*, vol. 36, no. 9, pp.11402- 11417.

2) Abdou, H., Pointon, J., 2011, "Credit scoring, statistical techniques and evaluation criteria: A review of the literature", *Intelligent Systems in Accounting, Finance and Management*, John Wiley & Sons, NJ, USA 18(2-3), pp.59–88.

3) Agresti, A., 2013, *Categorical Data Analysis*, Wiley & Sons, Florida, USA.

4) Anderson, R., 2007, *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*, Oxford University Press, Oxford, UK.

5) Baesens, B., Roesch, D. & Scheule, H., 2016, *Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS*, (Wiley and SAS Business Series), Wiley, New Jersey, USA.

6) DeLong, E.R., DeLong, D.M. & Clarke-Pearson, D.L., 1988, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach", *Biometrics*, International Biometric Society, vol. 44, no. 3, pp. 837-845.

7) Douglas, L. W., 2005, "Weight of Evidence: A review of concept and methods", *Risk Analysis*, Wiley, NJ, USA, vol. 25, no.6, pp.1545-1557.

8) Gouvêa, M.A., Gonçalves, E.B., 2021, "Credit risk analysis applying logistic regression, neural networks and genetic Algorithms Models", *International Journal of Advanced Engineering research and Science*, vol.8, no. 9, pp.198-209.

9) Good, I. J., 1950, *Probability and the Weighing of Evidence*, Charles Griffin & CO. LTD., London, UK.

10) Hand, D. J., Jacka, S. D., 1998, *Statistics in Finance*, Chapman and Hall, London, UK.

11) Hosmer, D. W. Jr., Lemeshow, S., & Sturdivant, R. X. ,2013, *Applied Logistic Regression*, Wiley, NJ, USA.

12) Lin, A.Z., Hsieh, T-Y., 2014, "Expanding the use of Weight of Evidence and information value to continuous dependent variables for variable reduction and scorecard development", South East SAS Users Group Conference (SESUG2014). https://www.lexjansen.com/sesug/2014/SD-20.pdf, (14/8/2023).

13) Siddiqi, N., 2006, *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*, John Wiley & Sons, New York, NY, USA.

14) Thomas, L.C., Crook, C.A & Edelman, D.J., 2002, *Credit Scoring and Its Applications*, Society for Industrial and Applied Mathematics, Philadelphia, USA.

15) Zeng, G., 2014, "A necessary condition for a good binning algorithm in credit scoring," *Applied Mathematical Sciences*, vol. 8, no. 65, pp.3229–3242.

16) https://www.myfico.com/credit-education/what-is-a-fico-score (14/8/2023)

17) https://www.myfico.com/credit-education/whats-in-your-credit-score (14/8/2023)

18) https://www.myfico.com/credit-education/fico-scores-vs-credit-scores (14/8/2023)

# Appendix 1

### Software -RStudio, R version 4.2.2

In the following segment of the research: Exploration of Data, Data Analysis, modelling logistic regression equation, Calculation of ROC-AUC values of the obtained models, Performance and Demonstration of Wald test for each model parameter, and Hosmer-Lemeshow test for obtaining the goodness of fit of the model.

### Code:

```
library(readxl)
data = read_excel('C:/Users/user/Desktop/Thesis/stats_voor2_Ahmadov_3900_truncated.xlsx')
names(data)
nrow(data)
unique(data$Status)
table(data$Sex)
table(data$Age)
table(data$Region)
data$Region[data['Region'] =="Võrumaa" ] = '_other'
data$Region[data['Region'] =="Lääne-Virumaa" ] = '_other'
data$Region[data['Region'] =="Pärnumaa" ] = '_other'
data$Region[data['Region'] =="Raplamaa" ] = '_other'
data$Region[data['Region'] =="Põlvamaa" ] = '_other'
data$Region[data['Region'] =="Järvamaa" ] = '_other'
data$Region[data['Region'] =="Viljandimaa" ] = '_other'
data$Region[data['Region'] =="Jõgevamaa" ] = '_other'
data$Region[data['Region'] =="Hiiumaa" ] = '_other'
data$Region[data['Region'] =="Läänemaa" ] = '_other'
data$Region[data['Region'] =="Valgamaa" ] = '_other'
data$Region[data['Region'] =="Saaremaa" ] = '_other'
table(data$Language)
is.numeric(data$Sum)
data$Sum = as.numeric(data$Sum)
is.numeric(data$Period)
is.numeric(data$Income)
is.numeric(data$Outcome)
table(data$Family)
table(data$Education)
```

```
table(data$WorkExperience)
is.numeric(data$Children)
is.numeric(data$Estate)
data <- data[data[,'PaimentAlertsTotal'] != 'NA', ]
data$PaimentAlertsTotal = as.numeric(data$PaimentAlertsTotal)
unique(data$PaimentAlertsTotal)
is.numeric(data$PaimentAlertsActive)
data$PaimentAlertsActive = as.numeric(data$PaimentAlertsActive)
data <- data[complete.cases(data$PaimentAlertsActive),]
sum(is.na(data['PaimentAlertsActive']))
is.numeric(data$PaimentAlertsClosed)
data$PaimentAlertsClosed = as.numeric(data$PaimentAlertsClosed)
data <- data[complete.cases(data$PaimentAlertsClosed),]
sum(is.na(data['PaimentAlertsClosed']))
data$Family = factor(data$Family)
levels(data$Family)
data <- within(data, Family <- relevel(Family, ref = 'Vallaline'))
levels(data$Family)
data$Education = factor(data$Education)
levels(data$Education)
data <- within(data, Education <- relevel(Education, ref = 'keskharidus'))
levels(data$Education)
data$WorkExperience = factor(data$WorkExperience)
levels(data$WorkExperience)
data <- within(data, WorkExperience <- relevel(WorkExperience, ref = 'Rohkem kui aasta'))
levels(data$WorkExperience)
data$Sex = factor(data$Sex)
levels(data$Sex)
data$Region = factor(data$Region)
levels(data$Region)
data <- within(data, Region <- relevel(Region, ref = 'Harjumaa'))
levels(data$Region)
data$Language = factor(data$Language)
levels(data$Language)
library(generalhoslem)
library(pROC)
library(aod)
```

```
model0=
glm(Status~Sex+Age+Region+Language+Sum+log(Period)+Outcome+Income+Children+Estate+PaimentAlertsActive+PaimentAlertsClosed+Family+Education+WorkExperience,data=data, family=binomial)

summary(model0)

logitgof(data$Status,fitted(model0))

HL_0=logitgof(data$Status,fitted(model0))

prob <- predict(model0, data, type = "response")

roc_curve <- roc(data$Status, prob)

plot(roc_curve, main = "ROC Curve")

auc(roc_curve)

AUC_0 = auc(roc_curve)

auc_ci_0 <- ci.auc(roc_curve)

se_0 <- sqrt(AUC_0 * (1 - AUC_0) / length(data$Status))

wald1 = wald.test(Sigma = vcov(model0), b = coef(model0), Terms = 1)

wald1

wald2 = wald.test(Sigma = vcov(model0), b = coef(model0), Terms = 2)

wald2

.

.

.

wald27 = wald.test(Sigma = vcov(model0), b = coef(model0), Terms = 27)

wald27

model1 =
glm(Status~Sex+Age+Region+Sum+log(Period)+Outcome+Children+Estate+PaimentAlertsActive+PaimentAlertsClosed+Education+WorkExperience,data=data, family=binomial)

summary(model1)

wald1 = wald.test(Sigma = vcov(model1), b = coef(model1), Terms = 1)

wald1

wald2 = wald.test(Sigma = vcov(model1), b = coef(model1), Terms = 2)

wald2

wald3 = wald.test(Sigma = vcov(model1), b = coef(model1), Terms = 3)

wald3

.

.

.

wald21 = wald.test(Sigma = vcov(model1), b = coef(model1), Terms = 21)

wald21

logitgof(data$Status,fitted(model1))

HL_1=logitgof(data$Status,fitted(model1))
```

```
prob <- predict(model1, data, type = "response")
roc_curve <- roc(data$Status, prob)
plot(roc_curve, main = "ROC Curve")
auc(roc_curve)
AUC_1 = auc(roc_curve)
auc_ci_1 <- ci.auc(roc_curve)
se_1 <- sqrt(AUC_1 * (1 - AUC_1) / length(data$Status))
```

# Appendix 2

### Software -RStudio, R version 4.2.2

In the following segment of the research: Application of correct binning strategy and Weight of Evidence transformation of non-binary variables

### Code:

```
data['Age_BIN'] = ifelse(data$Age <= 25 , "18-25",
                  ifelse (data$Age <= 30, "25-30",
                      ifelse (data$Age <= 35, "30-35",
                          ifelse(data$Age <= 40, "35-40",
                              ifelse(data$Age <= 45, "40-45",
                                  ifelse(data$Age <= 50, "45-50",
                                      ifelse(data$Age <= 55, "50-55",
                                          ifelse(data$Age <= 60, "55-60","60-70")))))))


table(data$Age_BIN,data$Status)


NR_18_25 = 196/(1160) ; R_18_25 = 236/(2823) ; WOE_18_25 = log(R_18_25/NR_18_25)
NR_25_30 = 218/(1160) ; R_25_30 = 351/(2823) ; WOE_25_30 = log(R_25_30/NR_25_30)
NR_30_35 = 159/(1160) ; R_30_35 = 384/(2823) ; WOE_30_35 = log(R_30_35/NR_30_35)
NR_35_40 = 167/(1160) ; R_35_40 = 401/(2823) ; WOE_35_40 = log(R_35_40/NR_35_40)
NR_40_45 = 147/(1160) ; R_40_45 = 350/(2823) ; WOE_40_45 = log(R_40_45/NR_40_45)
NR_45_50 = 108/(1160) ; R_45_50 = 310/(2823) ; WOE_45_50 = log(R_45_50/NR_45_50)
NR_50_55 = 71/(1160) ; R_50_55 = 297/(2823) ; WOE_50_55 = log(R_50_55/NR_50_55)
NR_55_60 = 56/(1160) ; R_55_60 = 260/(2823) ; WOE_55_60 = log(R_55_60/NR_55_60)
NR_60_70 = 38/(1160) ; R_60_70 = 234/(2823) ; WOE_60_70 = log(R_60_70/NR_60_70)
data['WOE_Age'] = ifelse(data$Age_BIN == '18-25' , WOE_18_25,
                  ifelse (data$Age_BIN == '25-30', WOE_25_30,
                      ifelse (data$Age_BIN == '30-35', WOE_30_35,
                          ifelse(data$Age_BIN == '35-40', WOE_35_40,
                              ifelse(data$Age_BIN == '40-45', WOE_40_45,
                                  ifelse(data$Age_BIN == '45-50', WOE_45_50,
                                      ifelse(data$Age_BIN == '50-55', WOE_50_55,
                                          ifelse(data$Age_BIN == '55-60', WOE_55_60, WOE_60_70)))))))
```

```
table(data$WOE_Age,data$Age_BIN)
table(data$Region,data$Status)
NR_h = 573/(1160) ; R_h = 1628/(2823) ; WOE_h = log(R_h/NR_h)
NR_o = 283/(1160) ; R_o = 515/(2823) ; WOE_o = log(R_o/NR_o)
NR_i = 186/(1160) ; R_i = 416/(2823) ; WOE_i = log(R_i/NR_i)
NR_t = 118/(1160) ; R_t = 264/(2823) ; WOE_t = log(R_t/NR_t)

data['WOE_Region'] = ifelse(data$Region == 'Harjumaa' , WOE_h,
                      ifelse (data$Region == '_other', WOE_o,
                          ifelse (data$Region == 'Ida-Virumaa', WOE_i,WOE_t)))
table(data$WOE_Region,data$Status)
data['Sum_BIN'] = ifelse(data$Sum <= 100 , "0-100",
                  ifelse (data$Sum <= 200, "100-200",
                      ifelse (data$Sum <= 300, "200-300",
                          ifelse(data$Sum <= 400, "300-400",
                              ifelse(data$Sum <= 500, "400-500",
                                  ifelse(data$Sum <= 800, "500-800","800-2000"
                                  ))))))
table(data$Sum_BIN,data$Status)
NR_0_100 = 156/(1160) ; R_0_100 = 488/(2823) ; WOE_0_100 = log(R_0_100/NR_0_100)
NR_0_200 = 284/(1160) ; R_0_200 = 620/(2823) ; WOE_0_200 = log(R_0_200/NR_0_200)
NR_0_300 = 346/(1160) ; R_0_300 = 778/(2823) ; WOE_0_300 = log(R_0_300/NR_0_300)
NR_0_400 = 113/(1160) ; R_0_400 = 256/(2823) ; WOE_0_400 = log(R_0_400/NR_0_400)
NR_0_500 = 142/(1160) ; R_0_500 = 331/(2823) ; WOE_0_500 = log(R_0_500/NR_0_500)
NR_0_800 = 74/(1160) ; R_0_800 = 190/(2823) ; WOE_0_800 = log(R_0_800/NR_0_800)
NR_0_2k = 45/(1160) ; R_0_2k = 160/(2823) ; WOE_0_2k = log(R_0_2k/NR_0_2k)
data['WOE_Sum'] = ifelse(data$Sum_BIN == '0-100' , WOE_0_100,
                  ifelse (data$Sum_BIN == '100-200', WOE_0_200,
                      ifelse (data$Sum_BIN == '200-300', WOE_0_300,
                          ifelse(data$Sum_BIN == '300-400', WOE_0_400,
                              ifelse(data$Sum_BIN == '400-500', WOE_0_500,
                                  ifelse(data$Sum_BIN == '500-800', WOE_0_800, WOE_0_2k))))))
table(data$WOE_Sum,data$Status)

hist(log(data$Period),breaks=50)
data['Period_BIN'] = ifelse(log(data$Period) <= 3.4 , "3.4",
```

```
                    ifelse (log(data$Period) <= 3.5, "3.5",
                        ifelse (log(data$Period) <= 4.1, "4.1",
                            ifelse(log(data$Period) <= 4.5, "4.5",
                                ifelse(log(data$Period) <= 5.1, "5.1",
                                    ifelse(log(data$Period) <= 5.8, "5.8","5.8+"
                                    ))))))
table(data$Period_BIN,data$Status)
NR3.4 = 24/(1160) ; R3.4 = 260/(2823) ; WOE3.4 = log(R3.4/NR3.4)
NR3.5 = 369/(1160) ; R3.5 = 1096/(2823) ; WOE3.5 = log(R3.5/NR3.5)
NR4.1 = 208/(1160) ; R4.1 = 323/(2823) ; WOE4.1 = log(R4.1/NR4.1)
NR4.5 = 176/(1160) ; R4.5 = 267/(2823) ; WOE4.5 = log(R4.5/NR4.5)
NR5.1 = 93/(1160) ; R5.1 = 173/(2823) ; WOE5.1 = log(R5.1/NR5.1)
NR5.8 = 169/(1160) ; R5.8 = 340/(2823) ; WOE5.8 = log(R5.8/NR5.8)
NR5.8u = 121/(1160) ; R5.8u = 364/(2823) ; WOE5.8u = log(R5.8u/NR5.8u)
data['WOE_logPeriod'] = ifelse(data$Period_BIN == '3.4' , WOE3.4,
                        ifelse (data$Period_BIN == '3.5', WOE3.5,
                            ifelse (data$Period_BIN == '4.1', WOE4.1,
                                ifelse(data$Period_BIN == '4.5', WOE4.5,
                                    ifelse(data$Period_BIN == '5.1', WOE5.1,
                                        ifelse(data$Period_BIN == '5.8', WOE5.8, WOE5.8u))))))
table(data$WOE_logPeriod,data$Status)
hist(data$Outcome,breaks=50)
# <=100,<=200,<=300,<=400,<=500,>500
data['Outcome_BIN'] = ifelse(data$Outcome <= 100 , "100",
                        ifelse (data$Outcome <= 200 , "200",
                            ifelse (data$Outcome <= 300 , "300",
                                ifelse(data$Outcome <= 400 , "400",
                                    ifelse(data$Outcome <= 500 , "500","500+"
                                    )))))
table(data$Outcome_BIN,data$Status)
NR100 = 141/(1160) ; R100 = 366/(2823) ; WOE100 = log(R100/NR100)
NR200 = 358/(1160) ; R200 = 703/(2823) ; WOE200 = log(R200/NR200)
NR300 = 265/(1160) ; R300 = 602/(2823) ; WOE300 = log(R300/NR300)
NR400 = 162/(1160) ; R400 = 451/(2823) ; WOE400 = log(R400/NR400)
NR500 = 105/(1160) ; R500 = 302/(2823) ; WOE500 = log(R500/NR500)
NR500u = 129/(1160) ; R500u = 399/(2823) ; WOE500u = log(R500u/NR500u)
data['WOE_Outcome'] = ifelse(data$Outcome_BIN == '100' , WOE100,
```

```
                    ifelse (data$Outcome_BIN == '200', WOE200,
                        ifelse (data$Outcome_BIN == '300', WOE300,
                            ifelse(data$Outcome_BIN == '400', WOE400,
                                ifelse(data$Outcome_BIN == '500', WOE500,WOE500u)))))
table(data$WOE_Outcome,data$Status)
hist(data$Children,breaks=50)
data['Children_BIN'] = ifelse(data$Children == 0 , "0",
                     ifelse (data$Children == 1 , "1","1+"))
table(data$Children_BIN,data$Status)

NR0 = 691/(1160) ; R0 = 1774/(2823) ; WOE0 = log(R0/NR0)
NR1 = 266/(1160) ; R1 = 630/(2823) ; WOE1 = log(R1/NR1)
NR1u = 203/(1160) ; R1u = 419/(2823) ; WOE1u = log(R1u/NR1u)
data['WOE_Children'] = ifelse(data$Children_BIN == '1' , WOE1,
                     ifelse (data$Children_BIN == '0', WOE0,WOE1u))
table(data$WOE_Children,data$Status)
hist(data$Estate,breaks=50)
data['Estate_BIN'] = ifelse(data$Estate == 0 , "0",
                    ifelse (data$Estate == 1 , "1","1+"))
table(data$Estate_BIN,data$Status)
NR0 = 838/(1160) ; R0 = 1344/(2823) ; WOE0 = log(R0/NR0)
NR1 = 268/(1160) ; R1 = 1105/(2823) ; WOE1 = log(R1/NR1)
NR1u = 54/(1160) ; R1u = 374/(2823) ; WOE1u = log(R1u/NR1u)
data['WOE_Estate'] = ifelse(data$Estate_BIN == '1' , WOE1,
                    ifelse (data$Estate_BIN == '0', WOE0,WOE1u))
table(data$WOE_Estate,data$Status)
hist(data$PaimentAlertsActive,breaks=50)
data['PaimentAlertsActive_BIN'] = ifelse(data$PaimentAlertsActive == 0 , "no","yes")
table(data$PaimentAlertsActive_BIN,data$Status)
NR0 = 972/(1160) ; R0 = 2546/(2823) ; WOE0 = log(R0/NR0)
NR1 = 188/(1160) ; R1 = 277/(2823) ; WOE1 = log(R1/NR1)
data['WOE_PaimentAlertsActive'] = ifelse(data$PaimentAlertsActive_BIN == 'no',WOE0,WOE1)
table(data$WOE_PaimentAlertsActive,data$Status)
hist(data$PaimentAlertsClosed,breaks=50)
data['PaimentAlertsClosed_BIN'] = ifelse(data$PaimentAlertsClosed == 0 , "0",
                            ifelse(data$PaimentAlertsClosed == 2 , "2",
                                ifelse(data$PaimentAlertsClosed == 3 , "3",
```

```
                                    ifelse(data$PaimentAlertsClosed == 1 , "1","m"))))
table(data$PaimentAlertsClosed_BIN,data$Status)
NR0 = 588/(1160) ; R0 = 1846/(2823) ; WOE0 = log(R0/NR0)
NR1 = 243/(1160) ; R1 = 396/(2823) ; WOE1 = log(R1/NR1)
NR2 = 136/(1160) ; R2 = 216/(2823) ; WOE2 = log(R2/NR2)
NR3 = 76/(1160) ; R3 = 128/(2823) ; WOE3 = log(R3/NR3)
NR3u = 117/(1160) ; R3u = 237/(2823) ; WOE3u = log(R3u/NR3u)

data['WOE_PaimentAlertsClosed'] = ifelse(data$PaimentAlertsClosed == '0' , WOE0,
                                  ifelse(data$PaimentAlertsClosed == '2' , WOE2,
                                      ifelse(data$PaimentAlertsClosed == '3' , WOE3,
                                          ifelse(data$PaimentAlertsClosed == '1' , WOE1,WOE3u))))
table(data$WOE_PaimentAlertsClosed,data$Status)
table(data$Family,data$Status)
NR_v = 416/(1160) ; R_v = 805/(2823) ; WOE_v = log(R_v/NR_v)
NR_a = 284/(1160) ; R_a = 884/(2823) ; WOE_a = log(R_a/NR_a)
NR_la = 90/(1160) ; R_la = 329/(2823) ; WOE_la = log(R_la/NR_la)
NR_le = 26/(1160) ; R_le = 101/(2823) ; WOE_le = log(R_le/NR_le)
NR_va = 344/(1160) ; R_va = 704/(2823) ; WOE_va = log(R_va/NR_va)
data['WOE_Family'] = ifelse(data$Family=='Abielus' , WOE_a,
                        ifelse(data$Family=='Vallaline',WOE_v,
                            ifelse(data$Family=='Vabaabielus',WOE_va,
                                ifelse(data$Family=='Lesk',WOE_le,WOE_la))))
table(data$WOE_Family,data$Status)
table(data$Education,data$Status)
NR_ke = 467/(1160) ; R_ke = 1081/(2823) ; WOE_ke = log(R_ke/NR_ke)
NR_al = 11/(1160) ; R_al = 20/(2823) ; WOE_al = log(R_al/NR_al)
NR_ei = 17/(1160) ; R_ei = 32/(2823) ; WOE_ei = log(R_ei/NR_ei)
NR_ko = 133/(1160) ; R_ko = 593/(2823) ; WOE_ko = log(R_ko/NR_ko)
NR_ku = 331/(1160) ; R_ku = 863/(2823) ; WOE_ku = log(R_ku/NR_ku)
NR_po = 201/(1160) ; R_po = 234/(2823) ; WOE_po = log(R_po/NR_po)
data['WOE_Education'] = ifelse(data$Education=='keskharidus' , WOE_ke,
                          ifelse(data$Education=='algharidus',WOE_al,
                              ifelse(data$Education=='kõrgharidus',WOE_ko,
                                  ifelse(data$Education=='kutseharidus',WOE_ku,
                                      ifelse(data$Education=='ei ole',WOE_ei,WOE_po)))))
table(data$WOE_Education,data$Status)
```

```
table(data$WorkExperience,data$Status)
NR_ro = 801/(1160) ; R_ro = 2248/(2823) ; WOE_ro = log(R_ro/NR_ro)
NR_ka = 37/(1160) ; R_ka = 64/(2823) ; WOE_ka = log(R_ka/NR_ka)
NR_ku = 275/(1160) ; R_ku = 381/(2823) ; WOE_ku = log(R_ku/NR_ku)
NR_to = 47/(1160) ; R_to = 130/(2823) ; WOE_to = log(R_to/NR_to)
data['WOE_WorkExperience'] = ifelse(data$WorkExperience == 'Rohkem kui aasta' , WOE_ro,
                        ifelse (data$WorkExperience == 'Katseaeg', WOE_ka,
                            ifelse (data$WorkExperience == 'Kuni aasta', WOE_ku,WOE_to)))
table(data$WOE_WorkExperience,data$Status)
hist(data$Income,breaks=50)
# <=400,<=600,<=800,<=1000,<=1200,>1200
data['Income_BIN'] = ifelse(data$Income <= 400 , "400",
                    ifelse (data$Income <= 600 , "600",
                        ifelse (data$Income <= 800 , "800",
                            ifelse(data$Income <= 1000 , "1000",
                                ifelse(data$Income <= 1200 , "1200","1200+"
                                )))))
table(data$Income_BIN,data$Status)
NR400 = 159/(1160) ; R400 = 389/(476+206+330+389+735+687) ; WOE400 = log(R400/NR400)
NR600 = 344/(1160) ; R600 = 735/(476+206+330+389+735+687) ; WOE600 = log(R600/NR600)
NR800 = 309/(1160) ; R800 = 687/(476+206+330+389+735+687) ; WOE800 = log(R800/NR800)
NR1000 = 168/(1160) ; R1000 = 476/(476+206+330+389+735+687) ; WOE1000 = log(R1000/NR1000)
NR1200 = 72/(1160) ; R1200 = 206/(476+206+330+389+735+687) ; WOE1200 = log(R1200/NR1200)
NR1200u = 108/(1160) ; R1200u = 330/(476+206+330+389+735+687) ; WOE1200u = log(R1200u/NR1200u)
data['WOE_Income'] = ifelse(data$Income_BIN == '400' , WOE400,
                    ifelse (data$Income_BIN == '600', WOE600,
                        ifelse (data$Income_BIN == '800', WOE800,
                            ifelse(data$Income_BIN == '1000', WOE1000,
                                ifelse(data$Income_BIN == '1200', WOE1200,WOE1200u)))))
table(data$WOE_Income,data$Status)
```

# Appendix 3

**Software -RStudio, R version 4.2.2**

In the following segment of the research: Modeling logistic regression equation with WoE transformed variables. Calculation of ROC-AUC values of the obtained models, Performance and Demonstration of Wald test for each model parameter, and Hosmer-Lemeshow test for obtaining the goodness of fit of the model. Performance of DeLong test between models with WoE transformed variables and models with original variables.

**Code:**

```
model2=
glm(Status~Sex+WOE_Age+WOE_Region+Language+WOE_Sum+WOE_logPeriod+WOE_Income+WOE_Outco
me+WOE_Children+WOE_Estate+


WOE_PaimentAlertsActive+WOE_PaimentAlertsClosed+WOE_Family+WOE_Education+WOE_WorkExperience
,data=data, family=binomial)

summary(model2)



wald1 = wald.test(Sigma = vcov(model2), b = coef(model2), Terms = 1)

wald1

wald2 = wald.test(Sigma = vcov(model2), b = coef(model2), Terms = 2)

wald2

wald3 = wald.test(Sigma = vcov(model2), b = coef(model2), Terms = 3)

wald3

.

.

.

wald16 = wald.test(Sigma = vcov(model2), b = coef(model2), Terms = 16)

wald16

logitgof(data$Status,fitted(model2))

HL_2=logitgof(data$Status,fitted(model2))

prob <- predict(model2, data, type = "response")

roc_curve <- roc(data$Status, prob)

plot(roc_curve, main = "ROC Curve")

auc(roc_curve)

AUC_2 = auc(roc_curve)

auc_ci_2 <- ci.auc(roc_curve)

se_2 <- sqrt(AUC_2 * (1 - AUC_2) / length(data$Status))
```

```
model3=
glm(Status~Sex+WOE_Age+WOE_Region+WOE_Sum+WOE_logPeriod+WOE_Outcome+WOE_Children+WOE
_Estate+
WOE_PaimentAlertsActive+WOE_PaimentAlertsClosed+WOE_Education+WOE_WorkExperience,data=data,
family=binomial)

summary(model3)

logitgof(data$Status,fitted(model3))

HL_2woe=logitgof(data$Status,fitted(model3))

prob <- predict(model3, data, type = "response")

roc_curve <- roc(data$Status, prob)

plot(roc_curve, main = "ROC Curve")

auc(roc_curve)

AUC_3 = auc(roc_curve)

auc_ci_3 <- ci.auc(roc_curve)

se_3 <- sqrt(AUC_3 * (1 - AUC_3) / length(data$Status))

wald1 = wald.test(Sigma = vcov(model3), b = coef(model3), Terms = 1)

wald1

wald2 = wald.test(Sigma = vcov(model3), b = coef(model3), Terms = 2)

wald2

wald3 = wald.test(Sigma = vcov(model3), b = coef(model3), Terms = 3)

wald3

.

.

.

wald13 = wald.test(Sigma = vcov(model3), b = coef(model3), Terms = 13)

wald13

AUC_2 ;  AUC_0

AUC_3 ;  AUC_1

predictions0 = predict(model0,type='response')

predictions1 = predict(model2,type='response')

auc0 <- roc(response = data$Status, predictor = predictions0)$auc

auc2 <- roc(response = data$Status, predictor = predictions1)$auc

roc.test(response = data$Status,predictor1 = predictions0, predictor2 = predictions1, method = "delong")

predictions2 = predict(model1,type='response')

predictions3 = predict(model3,type='response')

auc1 <- roc(response = data$Status, predictor = predictions2)$auc

auc3 <- roc(response = data$Status, predictor = predictions3)$auc

roc.test(response = data$Status,predictor1 = predictions2, predictor2 = predictions3, method = "delong")
```

Non-exclusive licence to reproduce thesis and make thesis public

I, Suleyman Ahmadov

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, Weight of evidence methodology in logistic regression with application in credit scoring, supervised by Kalev Pärna.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Suleyman Ahmadov

14/8/2023