



UNIVERSITY OF WARSAW
FACULTY OF ECONOMIC SCIENCES

WORKING PAPERS

No. 16/2019 (301)

A COMPARISON OF CREDIT SCORING TECHNIQUES IN PEER-TO-PEER LENDING

ANETA DZIK-WALCZAK
MATEUSZ HEBA

WARSAW 2019



A comparison of credit scoring techniques in Peer-to-Peer lending

Aneta Dzik-Walczak, Mateusz Heba*

University of Warsaw, Faculty of Economic Sciences

** corresponding author: mheba@wne.uw.edu.pl*

Abstract: Credit scoring has become an important issue as the competition among financial institutions becomes very intense and even a slight improvement in accuracy of prediction might translate into significant savings. Financial institutions are seeking optimal strategies through the help of credit scoring models. Therefore credit scoring tools are widely studied. As a result different parametric statistical methods, non-parametric statistical tools and soft-computing approaches have been developed in order to increase the accuracy of credit scoring models.

In this paper different approaches to classify customers as those who pay back loan and those who default on a loan will be employed. The purpose of this study is to explore the performance of two credit scoring techniques, the logistic regression model and neural networks. In order to evaluate the feasibility and effectiveness of these methods analysis is performed on Lending Club data. Peer-to-Peer lending, also called social lending are investigated. On the basis of the results, we can conclude that logistic regression model can provide better performance than neural nets.

Keywords: credit scoring, credit risk, Lending Club, logistic regression, neural nets, peer-to-peer lending

JEL codes: G21; G32

1. Introduction

Credit risk is an inseparable aspect of lending. An important issue for financial institution is achievement the lowest fraction of bad loans by minimizing the asymmetry information between the lender and the borrower. Credit risk decisions are key determinants for the success of financial institutions due to very high costs of wrong decisions (Lahsasna *et al.*, 2010). Credit risk assessment is crucial element of credit risk management and states the basis for loan decision-making (Wu *et al.*, 2010). Due to the significance of credit risk, a number of tools to improve risk assessment and increase the prediction accuracy have been proposed. The credit scoring aim is to classify customers as good customers, the customers who paid back loan and bad customers, the customers who defaulted on a loan. Different parametric statistical methods, non-parametric statistical tools and soft-computing approaches have been developed. Artificial neural networks, genetic algorithms, genetic programming, support vector machines, and some hybrid models have been used to evaluate credit risk.

Peer-to-Peer (P2P) lending has emerged in recent years. P2P lending, also called social lending, consists in individuals lending money to other individuals. No financial institution is present in this process. In P2P lending individual investors bear the credit risk, instead of financial institutions. An electronic platform mediates between borrowers and lenders of loans, charging a fee for this service. Companies such as Lending Club channel loans between individuals.

The research question of this paper aims at comparing two credit score classification methods: logistic regression model and neural networks. In recent years, neural nets, are perceived as one of the best statistical techniques used in building the scoring models. The aim of the study was to verify the hypothesis that neural nets model is a better classifier compared to logistic regression. We use data from Lending Club, the biggest US P2P lending company. The sample analyzed contains 119160 observations loans from 2011 to 2013. Loans funded until 2013 are analyzed, as the status of later loans (defaulted or non-defaulted) is unknown because the maturity of these loans is 36 or 60 months.

The remainder of the paper is organized as follows. In Section 2, the literature review is presented. In Section 3, we describe the details of methodology. Section 4 reports the empirical results. Based on the results of our research, Section 5 draws conclusions.

2. Literature review

Credit scoring is a financial tool used in the risk assessment process, namely to manage and diversify the risk in investment portfolios. One of the main aim of credit scoring is to estimate risk of loans. Credit scoring models allow to assign loan customers to either good or bad category with respect to their characteristics. Credit scoring models are widely used by financial institutions to determine if loan customers belong to either a good applicant group or a bad applicant group. Credit scoring tools are used to evaluate new consumer loans as well as to analyze existing loans. Credit scoring models allow for achieving the benefit from reducing the cost of credit analysis, enable faster credit decision, diminish possible risk (Lee *et al.*, 2002; West, 2000). Scoring tasks are then related to classification analysis. Since an improvement of such classification might translate into significant savings, the accuracy of different credit scoring tools should be analyzed and compared.

In order to increase the accuracy of credit scoring model, numerous methods have been developed. Both parametric statistical methods (e.g. discriminant analysis, logistic regression) and non-parametric statistical methods (e.g. decision trees) are used. In recent years, many novel approaches such as artificial neural networks, rough sets, or decision trees have been proposed to improve credit scoring models.

Based on literature review it can be concluded, there is no overall best method used in building credit scoring models. The authors state that many factors such as the data structure, explanatory variables, criterion of comparison are relevant when selecting the best technique of the classification (e.g. Hand and Henley, 1997).

Abdou *at al.* (2008) find that the lowest misclassification cost criterion leads to net search selecting multi-layer feed-forward net with five nodes. However, probabilistic neural nets provide the highest average correct classification rate. Tsai *et al.* (2009) show that DEA-DA (Data Envelopment Analysis-Discriminant Analysis) and neural nets are possessed better predicting capability than probit analysis and logistic regression. Neural nets were picked as the optimal predicting model also by Yeh and Lien (2009). Kočenda and Vojtek (2011) detected that both credit risk models based on logistic regression and regression trees are comparably efficient. Wang *et al.* (2011) explored logistic regression analysis, decision tree, artificial neural network, support vector machine. Bagging decision tree got the best performance in terms of accuracy, type I error and type II error. Akkoç (2012) proposes a three stage hybrid adaptive neuro fuzzy inference system credit scoring model which performs better than the linear discriminant analysis, logistic regression and artificial neural network in terms of average correct classification rate and estimated misclassification cost. Bekhet and Eletter (2014)

indicate that in terms of the overall accuracy rate the logistic regression performs better than the radial basis function model. However, the radial basis function is superior in identifying customers who may default. Tsai *et al.* (2014) tried to reduce exposure to loan defaults and exceed Lending Club's return at the same risk level. They used four algorithms: naive Bayes, random forest, support vector machines and modified logistic regression (penalty if the classifier incorrectly classifies a defaulted loan as a good loan). The highest precision was obtained for modified logistic regression on two-dimension principal component analysis data. Chang, Kim and Kondo (2015) compared performance of logistic regression, naive Bayes, support vector machine (with different commonly used kernels: linear, polynomial, Gaussian radial basis function, and sigmoid). To indicate the performance of the model they used sensitivity, specificity, the G-mean, completion of performance metrics, accuracy and precision. They found that Naive Bayes with Gaussian performs the best with default prediction (80.1% sensitivity). Malekipirbazari and Aksakalli (2015) declared a random forest based classification method to be the best classifier compared to FICO (a publicly traded corporation, produces scoring models that are most commonly used and distributed by TransUnion, Equifax, and Experian), logistic regression, support vector machines. Random forests had the highest accuracy rate, the highest AUC and the lowest RMSE. Fitzpatrick and Mues (2016) evaluate the performance of boosted regression trees, random forests, penalized linear and semi-parametric logistic regression models and find that boosted regression trees outperform logistic regression. Imtiaz and Brimicombe (2017) conclude that artificial neural network is better alternative than decision tree and logistic regression when data availability is high in dataset.

The summary of methods compared in empirical articles is shown in the table 1. Logistic regression, neural networks, decision trees, discrimination analysis and support vector machines are the most frequently compared techniques. Neural networks outperform other methods the most often.

Table 1. Credit scoring techniques used in various studies

Article/Technique	Logistic regression	Neural networks	Decision trees	Discriminant analysis	Support vector machines	Random forests	Naive Bayes	K-nearest neighbors	Survival time analysis	Proportional hazards model	Probit model	Generalized additive model	Graphic analysis with discriminant analysis	ANFIS system
Abdou, Pointon, El-Masry (2008)	•	• I		•							•			
Ince, Aktan (2009)	•	• I	•	•										
Tsai, Lin SP., Cheng, Lin YP. (2009)	•	• I											• I	
Yeh, Lien (2009)	•	• I	•	•			•	•						
Kočenda, Vojtek (2011)	• I		•											
Wang, Hao, Ma, Jiang (2011)	•	•	• I		•									
Akkoç (2012)	•	•		•										• I
Bekhet, Eletter (2014)	• I	•												
Abdou, Dongmo Tsafackc, Ntimd, Baker (2016)	•	• I	•											
Fritzpatrick, Mues (2016)	•		• I			•						•		
Imtiaz, Brimicombe (2017)	•	• I ⁱ	• I ^{bi}											
Serrano-Cinca, Gutiérrez-Nieto, López-Palacios (2015)	•								•					
Emekter, Tu, Jirasakuldech, Lu (2015)	•									•				
Tsai, Ramiah, Singh (2014)	• I				•	•	•							
Chang, Kim, Kondo (2015)	•				•		• I							
Malekipirbazari, Aksakalli (2015)	•				•	• I		•						
Legend: • – technique used in a study I – technique stated as the best in a study (I ⁱ – with data imputation, I ^{bi} – without data imputation)														

Source: Own elaboration.

3. Methodology

In this section, two statistical techniques used in credit scoring are described first. The first method is the logistic regression model. The second method is neural nets.

3.1. Logistic regression

Logistic regression is a modelling technique in which the probability of a binary outcome is related to a set of predictor variables. Logistic regression is the most commonly used technique in the area of credit scoring - in particular when classifying clients. This model can be used, among others for analyzing the probability of timely repayment of a loan, as well as determining the customer's belonging to one of two groups - reliable or unreliable borrowers (Matuszyk, 2018). Thus, the objective of a logistic regression model in credit scoring is to determine the conditional probability of a specific observation belonging to one of two groups (good and bad customers), given the values of the independent variables of that observation. Bad customers refers to the customers who defaulted on a loan and good customers refers to the customers who paid back loan.

The dependent variable in the logistic regression model is the binary variable, and the method of estimation is maximum likelihood. The scoring model in our research was estimated on categorized variables modified with the use of WOE (Weight of Evidence) transformation. WOE allows to transform a continuous independent variable into a set of bins based on similarity of dependent variable distribution i.e. number of events. Such approach solves the problems related to outliers, allows modeling of non-linear relationships using linear models, gives the opportunity to better interpret the relationships found in the data (Siddiqi, 2006). WOE allows measuring the diversity between the analyzed groups (e.g. borrowers paying their liabilities on time and borrowers not paying their payments) within a given category of the selected variable and is calculated as follows:

$$WOE_i = \ln \left(\frac{good_i}{bad_i} \right),$$

where: i – category of a variable, $good_i$ – the percentage of customers who repay loans on time for category i , bad_i – percentage of customers who do not pay their liabilities on time for category i .

WOE is used to calculate IV (Information Value) to select important variables in a predictive model:

$$IV = \sum_{i=1}^n (good_i - bad_i) \cdot WOE_i ,$$

If the IV statistic is less than 0.02, then the predictor is not useful for modeling (separating the Goods from the Bads). Values between 0.02 and 0.1, indicate a weak predictor. IV from 0.1 to 0.3 means a medium predictor. If IV ranges from 0.3 to 0.5, then the predictor has a strong relationship to the Goods/Bads odds ratio.

An alternative measure of the predictive power of variables is the Gini coefficient:

$$Gini = 1 - \sum_{i=1}^n ((cbad_i - cbad_{i-1}) \cdot (cgood_i - cgood_{i-1})),$$

where: n – number of categories, $cgood_i$ – the cumulative percentage of customers who repay loans on time for category i , $cbad_i$ – cumulative percentage of customers who do not pay their liabilities on time for category i .

3.2. Neural nets

The idea of neural nets comes from the structure of the human brain. In the human brain a neuron may send (receive) a signal to (from) other neurons. Similarly, neural nets consist of elements, each of which receive a number of inputs, and generate a single output. Neural nets composed of a number of simple nodes or neurons elements, which are connected together from either a single layer or multiple layers.

An important element of the activity of each neuron is the function responsible for the value of the output signal, which is called the activation function. It can take one of three forms: binary step, linear and non-linear (e.g. sigmoid). As a result an output value (node value) Y is given:

$$Y = g \left(\sum_{i=0}^{n-1} w_i y_i - \varphi_h \right),$$

where: Y – output signal value (node value), i – selected input signal, y_i – value of the selected input signal, w_i – weight of the selected input signal, $g(*)$ – activation function, φ_h – threshold activation level.

Neural nets have many advantages. One of the most important: there is no need to meet assumptions about statistical distribution of variables and error terms (Matuszczyk, 2018). However, neural nets have also some disadvantages, including: the possibility of a very long learning process, the possibility of instability of behavior in the learning process, the ability to end the action of finding the local minimum (without finding the optimum), poor network transparency – difficulties in interpretation and difficult to locate the cause of errors occurring.

We use a popular type of neural net - the multilayer perceptron (MLP) in our analysis. The MLP net is an unidirectional net, which usually has in its structure: an input layer, one (or two) hidden layers built of sigmoidal neurons, and an output layer built of sigmoid or linear neurons. The backpropagation algorithm is used in the learning process.

3.3. Assessing the quality of classification models

3.3.1. Confusion matrix

Very often it may not be sufficient to know how often the evaluated model is wrong. It may be more important to know how often it fails to predict correctly some specific outcome (Liu, 2002). The confusion matrix gives useful insight into the model's capability to predict particular group. The matrix includes four possible cases (assuming that there are only two categories - positive and negative): *True Positives (TP)* – the number of positive cases correctly classified into positive class, *False Negatives (FN)* – the number of positive cases incorrectly classified as negative, *False Positives (FP)* – the number of negative cases incorrectly classified as positive, *True Negatives (TN)* – the number of correctly classified negative cases to a negative class.

A number of performance measures can be derived from the confusion matrix, for example:

- *accuracy* the ratio of correctly classified cases to all cases $\frac{TP+TN}{TP+FN+FP+TN}$,
- *error ratio* the ratio of incorrectly classified cases to all cases $\frac{FP+FN}{TP+FN+FP+TN}$,
- *positive predictive value* (true positive rate) the ratio of cases correctly classified as positive to all positive cases $\frac{TP}{TP+FP}$,
- *negative predictive value* (false positive rate) the ratio of cases incorrectly classified as positive to all negative cases $\frac{FP}{TN+FP}$,
- *recall, sensitivity* (precision) the ratio of cases correctly classified as positive to all cases classified as positive $\frac{TP}{TP+FP}$,
- *specificity* the ratio of cases correctly classified as negative to all negative cases (the same as 1 – false positives rate) $\frac{TN}{TN+FP}$

Type I error rate is the rate of bad customers being categorized as good. A high type I error rate means that the institution is exposed to credit risk. Type II error rate (called also β) is the rate of good customers being categorized as bad. A high type II error rate means that the institution is exposed to commercial risk as a high β rate for a long time, means that institution

takes a long time restrictive credit granting policy, and may lose its share in the market (Boggess, 1967).

Given confusion matrix is calculated for stated cut off point which is a certain threshold value used to determine whether an observation belongs to a particular class.

3.3.2. ROC Curve

One convenient tool that facilitates the performance of a classification model at all classification thresholds is the ROC (Receiver Operating Characteristic) analysis. The ROC curve plots true positives rate vs. false positives rate at different classification cut off points. On the ROC curve, the x-axis is identified with the measure of unity minus the specificity, while the y-axis is the sensitivity. To draw a ROC curve for a scoring classifier confusion matrices for different cut off points are calculated. An reasonable scoring classifier should have their ROC curve entirely above the diagonal (the random model line – $y = x$ line, passing through (0,0) and (1,1)), which means the true positive rate should be always above the false positive rate.

3.3.3. AUC

In order to facilitate comparison of classification models, the area underneath the entire ROC curve can be calculated – AUC (*Area Under ROC Curve*). AUC provides an aggregate measure of performance across all possible classification thresholds. AUC ranges in value from 0 to 1. The ideal classifier has an AUC measure of 1 (predictions are 100% correct), while for the random classifier this indicator is 0.5. The larger the AUC measure, the better the classification model. AUC measures the quality of the model's predictions irrespective of what classification threshold is chosen.

3.3.4. Gini coefficient

Another measure of goodness of a binary classifier is Gini coefficient. Its calculation is based on mentioned above the area underneath the entire ROC curve (AUC). The Gini coefficient is a ratio between: (1) the area between the ROC curve and the random model line and (2) the top left triangle above the random model line.

$$Gini = \frac{AUC - 0.5}{0.5} = \frac{AUC}{0.5} - 1 = 2 \cdot AUC - 1 ,$$

where: AUC – the area underneath the entire ROC curve.

The higher the Gini coefficient, the better the model.

3.3.5. Kolmogorov-Smirnov statistic

Kolmogorov-Smirnov statistic is defined as the maximum difference between true positive rate (the probability that the model detects an actual positive as positive) and false positive rate (the probability that the model detects an actual negative as positive) obtained for different cut off points. Higher Kolmogorov-Smirnov statistic value is indicative of better model. It is defined as:

$$KS = \max_{a \in [L, H]} [abs(F_{m,bad}(a) - F_{n,good}(a))],$$

where: a – score that ranges from L to H , L – the minimum value of a given score, H – the maximum value of a given score, $F_{m,bad}(a)$ – the empirical cumulative distribution function of the scores of bad clients, $F_{n,good}(a)$ – the empirical cumulative distribution function of the scores of good clients.

4. Results

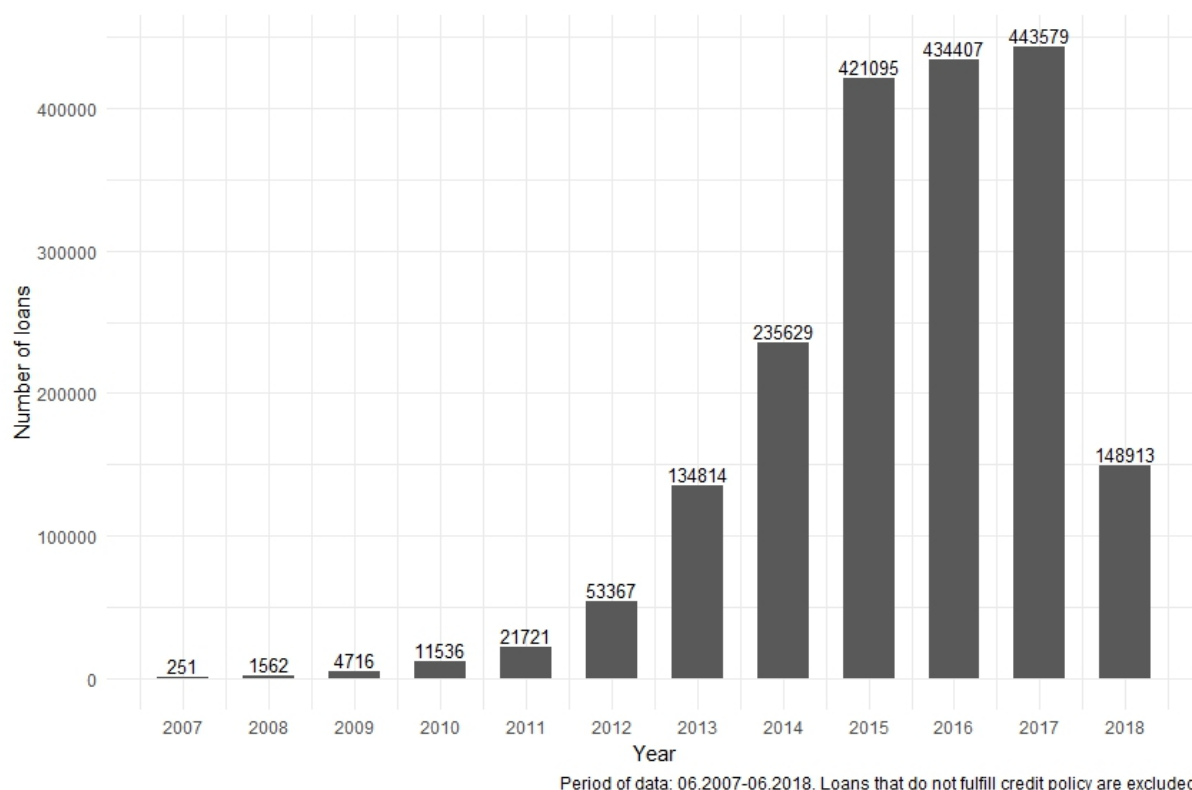
In this section results of empirical research are presented.

4.1. Data – source and preparation for modelling

Using data provided by financial institutions (banks) is very difficult due to different legal regulations and restrictions. Therefore, data that is used in this research was downloaded from the website of one of the largest USA peer-to-peer lending money company – Lending Club¹. This company provides data mainly for potential investors. However, it is publicly available. Data was downloaded on July 15, 2018. It is important information because Lending Club provides the most current credit portfolio data (data is provided every quarter). That is why some information about older loans was updated at the end of the second quarter of 2018. According to the comments available on the company's website, the update of older data includes essentially current loan status and recent repayment history. Every quarter Lending Club provides new part of the database that includes loans that are granted in this quarter. Several datasets were downloaded from the website that covered time period from the beginning of June 2007 to the end of June 2018. As a result of datasets merging one big dataset was obtained. It included information about 1911592 loans (without 2749 credits that does not satisfy internal credit policy conditions) with almost 28.5 billion dollars in total. Recently, the peer-to-peer lending has become more and more popular. This trend can be noticed e.g. based

¹ Lending Club, data: <https://www.lendingclub.com/info/download-data.action> [accessed 27.08.2018].

on the number of credits that were granted by Lending Club in different years what is shown in the figure 1.



Source: Own elaboration.

Figure 1. Number of credits granted by Lending Club in different years

One of the most important step during model building process is selection of data timeframe for model estimation. There are few rules of data timeframe selection. However, using Lending Club data has some limitations, namely lack of access to whole repayment history for loans. In some cases it is possible to use only the most current data. It makes analysis of immaturity credits difficult because of the fact that there is not the same data time-window in the context of credit defaults occurrence. That is why in this paper only matured loans are considered. Lending Club has two different types of credits in the context of maturity time: 36-month and 60-month. Thereupon, it is necessary to exclude data from last 5 years in available dataset in order to have only matured loans in sample. Thus, the end of the second quarter 2013 is upper boundary of time in the analysis. In credit scoring usually the 12-month timeframe is used in model development process, but the time window can be extended to 24-month period (Matuszyk, 2018). In this article 2-year timeframe is chosen and data from July 1, 2011 to July 30, 2013 is considered.

Next very important step in prediction model development process is defining positive and negative events that divides all observations into two separated groups. In this analysis it is

important to differentiate customers who repay loans in a timely manner and who do not do it on time. Definitions of good and bad events are based on variable that describes loan status. This variable has many possible values but two of them are used – *Fully Paid* loans that are interpreted as negative event (marked as 0) and *Charged Off* credits that are considered as positive event (marked as 1). Other loans statuses are filtered out. Data set which is used has 119160 observations and 151 variables.

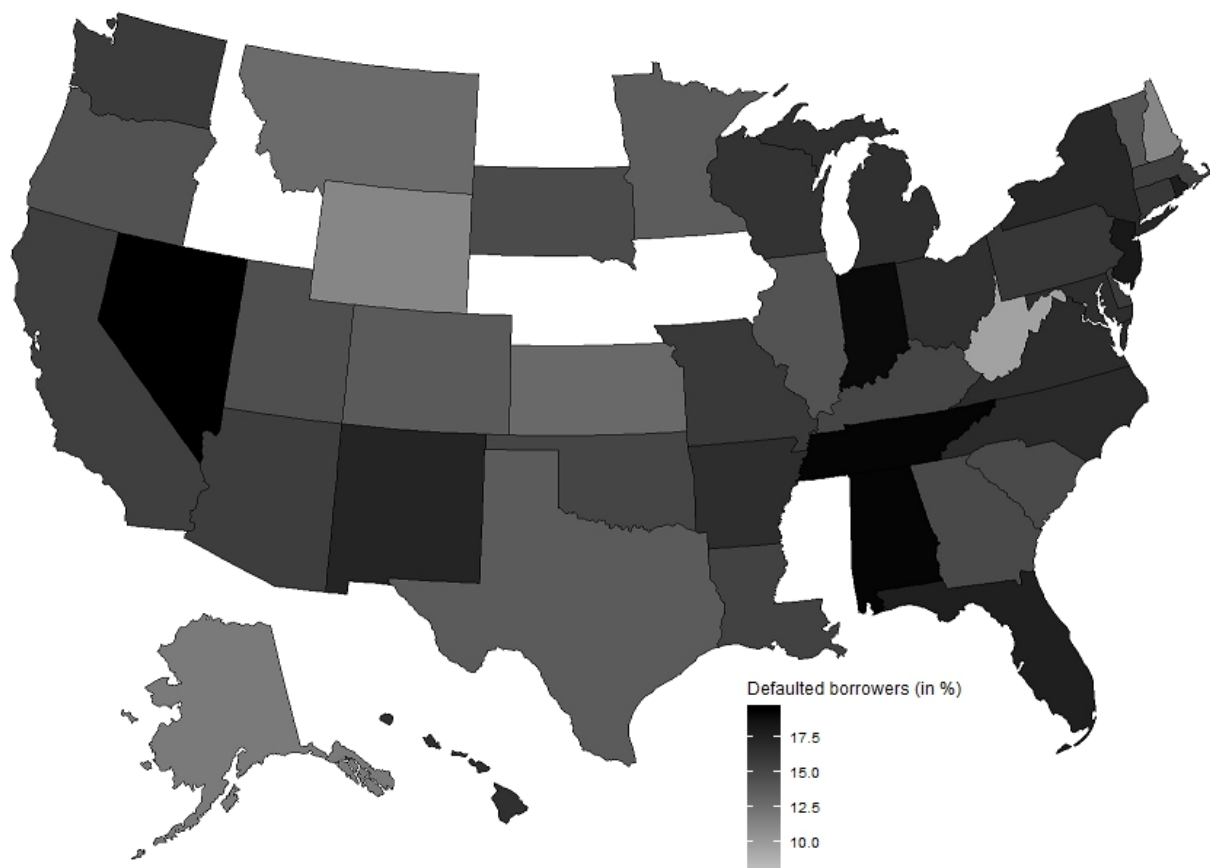
Posterior step of model building process includes preliminary data analysis which is essential in the context of big datasets. At the beginning variables with a lot of missing values are eliminated (at least 90% of missing values). It reduces 53 variables. Afterwards, technical variables are removed (links to websites, observation IDs and similar), characteristics that are hard to use in modeling process (e.g. unstructured job description delivered by borrower) and variables that are not known at the moment of credit application – there are not such an information in the document filled by borrower in (e.g. instalment value, loan risk class, interest rate, last payment amount). It reduces next 32 variables.

Some of remaining variables are hard to use in the context of modeling in their original form, e.g. comment about credit delivered by borrower or date of opening the first credit product by the borrower. Because of this fact, new variables are created (based on original variables): number of letters in credit comment (there is a suspicion that people who add very long comment as loan need motivation are more risky because they want to get the credit at all costs), number of years from the loan release (assuming that date of loan release is almost the same as date of submitting the credit application by a borrower). What is more, the new loan status variable is added as binary variable (accordingly with previous description) – fully paid credit is marked as 0 and default is marked as 1. Additionally, date of job start is transformed from string type to numeric one (0 means working period shorter than a year and 10 means working period that lasts at least 10 years). Unnecessary original variables are removed from dataset.

Thereafter, the problem of missing values are considered (for variables that in their original form have less than 90% of missing values). This problem relates to variables that store numerical information e.g. number of different products (particular types), number of months from payment delay with different time and amount thresholds. A lack of data in this context presents additional information. That is why missing values are replaced with value -1 in order to use this information in modelling process. It eliminates problem with missing values for 40 variables.

In the next step the variables filled with the same value for all observations are removed from the analysis. It reduces 3 variables. Finally, there are 63 variables (including 1 explained variable).

After analyzing distributions of different variables, it is necessary to transform some of them. For home ownership variable two categories – *other* and *none* – are combined into one group *other* because of the small number of observations in those two separate categories. The next variable that is modified is characteristic with state where the borrower lives because this variable has too many categories. That is why for all states the ratio of defaulted borrowers (in comparison to all borrowers) is calculated and then a few states with smallest and a few states with largest values of this ratio are selected, whereby particular states have to have at least 100 citizens with Lending Club credits in order to be finally selected. The distribution of this ratio is presented on the USA map in the figure 2.



Source: Own elaboration.

Figure 2. Ratio of defaulted borrowers by USA states²

² The Map is created in black and white colors. The white color is used for the states having no data or the states which the limitation condition is unfulfilled for.

Basing on calculated ratio, the new variable is created – the new characteristic for all the states assigns one of three classes: *Max DR* for the locations with the largest values of the defaulted borrowers ratio (states with codes: NV, AL, TN, IN, NJ), *Min DR* for the areas with smallest values of the ratio (states with codes: DC, WV, NH, WY, AK) and *No info* for other states. Unnecessary original variables are removed from dataset.

Some of numeric variables have only few distinct values (integers) – those of them that have less than 10 distinct values are considered as ordinal variable. What is more, because of a small number of observations in some categories of particular variables, some of groups are merged. The most often considering 3-group allocation is following: -1 (missing values), 0 i 1+ (more than 1). Such an transformation is used for 5 characteristics, e.g. number of accounts with any payment delinquency. However, in some cases other 3-group allocations are used, namely: 0, 1 i 2+ (more than 2) or 4-group allocation is considered: 0, 1, 2 i 3+ (more than 3). It is applied for 2 variables – number of public record bankruptcies and the number of credit inquiries in past 6 months by the borrower.

In the next step of model building process the dataset is limited in order to have only the most predictive variables. To do the variable selection, the *fine classing* procedure is applied. It is the binning technique – it bins the variable into many quantile groups (e.g. 10) and for each group following measures are calculated: the bad rate (shows proportion of observations having the value 1 for the target variable in comparison to all records in given group) and the WOE (*Weight of Evidence*). It is useful to calculate the variable's predictive power measures – Gini coefficient and IV (*Information Value*). Those indicators are very helpful in variable selection process. However, before the final variable selection, the correlation analysis is done. The Kendall's Tau correlation coefficient is used as measure of relationship between two different ordinal variables because dataset includes transformed variables using WOE approach (not original data) and consequently they are ordinal characteristics. If the correlation coefficient value between two explanatory variables is more than 0.5, the explanatory variable that is more correlated with target variable is kept in the dataset (the other one is eliminated from the dataset). After correlation analysis the final variable selection is done basing on Gini coefficient – variables that have value of this coefficient more than 5% are considered as the most predictive variables. According to this rule final variable set includes 17 characteristics – target variable and 16 explanatory variables. List of final characteristics (with information about Gini coefficient values) using in models estimation process is presented in the table 2.

Table 2. List of final characteristics using in models estimation process

Variable	Variable description	Gini coefficient
Target variable		
<i>Loan_status_n</i>	Loan status (modified original variable)	–
Explanatory variables		
<i>Term</i>	The number of payments (in months)	0.17113
<i>Acc_open_past_24mths</i>	The number of trades opened in past 24 months	0.11715
<i>Dti</i>	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested Lending Club loan, divided by the borrower's self-reported monthly income	0.11669
<i>Revol_util</i>	Revolving line utilization rate	0.10943
<i>Annual_inc</i>	The self-reported annual income provided by the borrower during registration	0.10906
<i>Percent_bc_gt_75</i>	Percentage of all bankcard accounts greater than 75% of limit	0.1046
<i>Verification_status</i>	Indicates if income was verified by Lending Club, not verified, or if the income source was verified	0.10199
<i>Inq_last_6mths</i>	The number of inquiries in past 6 months (modified original variable)	0.10078
<i>Loan_amnt</i>	The amount of the loan	0.09831
<i>Num_tl_op_past_12m</i>	The number of accounts opened in past 12 months	0.09365
<i>Purpose</i>	A category provided by the borrower for the loan request	0.09249
<i>Mo_sin_rcnt_tl</i>	Months since most recent account opened	0.08401
<i>Mths_since_recent_bc</i>	Months since most recent bankcard account opened	0.07545
<i>Num_rev_tl_bal_gt_0</i>	The number of revolving trades with balance greater than 0	0.07218
<i>Mort_acc</i>	The number of mortgage accounts	0.05926
<i>Earliest_cr_line_n</i>	The number of months from the date that the borrower's earliest reported credit line was opened (modified original variable)	0.05216

Source: Own calculations. Description of variables delivered by Lending Club.

In final dataset there are 5 categorical and 12 continuous variables. Values of categorical variables are following³:

- *Loan_status_n*: 0 (84.1%), 1 (15.9%).
- *Term*: 36 months (78%), 60 months (22%).
- *Verification_status*: verified (45.5%), not verified (34.4%), source verified (20.1%).
- *Inq_last_6mths*: 0 (50%), 1 (28.3%), 2 (13.7%), 3+ (8%).
- *Purpose*: debt consolidation (57.9%), credit card (21%), home improvement (5.8%), other (5.2%), major purchase (2.4%), small business (2.1%), car (1.5%), wedding (1.1%), medical (1%), house (0.7%), moving (0.7%), vacation (0.5%), renewable energy (0.1%).

Descriptive statistics of continuous variables are shown in the table 3.

³ Percentage values in brackets show distribution of particular variables categories.

Table 3. Descriptive statistics of continuous variables 4

Variable/statistic	Average	Standard deviation	Minimum	Median	Maximum	Kurtosis	Missing percentage
<i>Acc_open_past_24mths</i>	3.91	2.68	0.00	3.00	40.00	3.19	16.80
<i>Dti</i>	16.67	7.59	0.00	16.44	34.99	0.63	0.00
<i>Revol_util</i>	58.08	24.00	0.00	60.90	122.50	-0.57	0.10
<i>Annual_inc</i>	71546.84	55790.00	4800.00	61000.00	7141778.00	3875.78	0.00
<i>Percent_bc_gt_75</i>	34.64	27.53	0.00	33.00	94.00	-1.32	17.50
<i>Loan_amnt</i>	14046.85	8168.86	1000.00	12000.00	35000.00	-0.10	0.00
<i>Num_tl_op_past_12m</i>	1.82	1.57	0.00	2.00	25.00	4.77	33.80
<i>Mo_sin_rcnt_tl</i>	8.95	9.67	0.00	6.00	99.00	13.80	33.80
<i>Mths_since_recent_bc</i>	22.14	21.69	0.00	14.00	99.00	1.17	17.40
<i>Num_rev_tl_bal_gt_0</i>	5.75	2.96	0.00	5.00	37.00	2.36	33.80
<i>Mort_acc</i>	1.74	2.19	0.00	1.00	29.00	3.41	16.80
<i>Earliest_cr_line_n</i>	14.59	6.93	3.00	13.00	62.00	1.78	0.00

Source: Own calculations.

Additionally, the table 4 presents the results of independence Chi² test between the target variable and the categorical explanatory variables.

Table 4. Results of Chi2 test (target variable vs. categorical explanatory variables)

Variable	Chi ² statistics	p-value
<i>Term</i>	2719.3	0.0000
<i>Verification_status</i>	597.12	0.0000
<i>Inq_last_6mths</i>	585.46	0.0000
<i>Purpose</i>	598.11	0.0000

Source: Own calculations.

According to the results from table 4, for all the variables the null hypothesis about variable independence is rejected (p-value in each case is lower than the assumed 5% significant level), hence there are statistical significant dependents between target variable and particular explanatory variables.

In the next step dataset is randomly divided into training and validation subsets in proportion 70:30. In each subsets the proportion of defaults is similar as in the original dataset (before division). Training subset has 83412 observations and ca. 15.9234% defaulted credits. Validation subset has 35748 records and ca. 15.9226% defaulted loans.

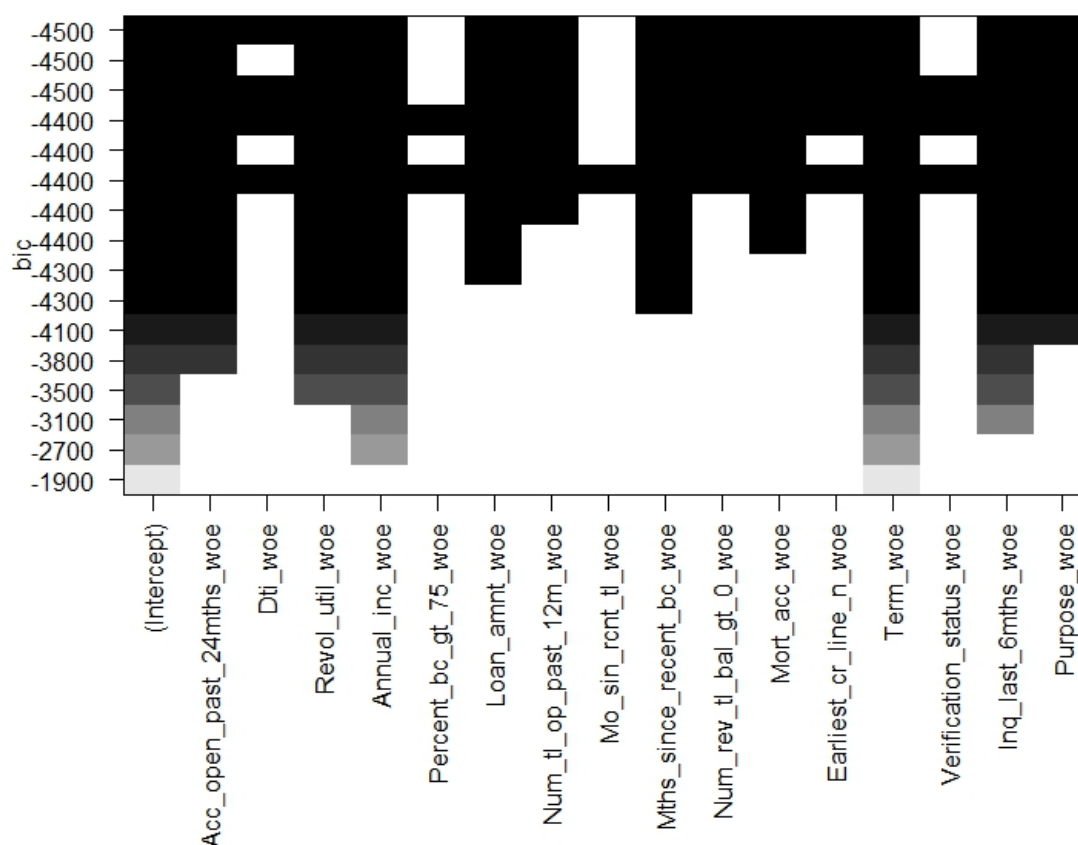
4.2. Logistic regression model

According to methodology description – scoring model that uses logistic regression model is based on variables after WOE transformation (instead of their original values). This

⁴ Descriptive statistics do not take into consideration introducing -1 values as missing values.

transformation is used in the *fine classing* procedure. However, in the process of logistic regression model estimation, variables after WOE transformation in the *coarse classing* procedure are used. Those two procedures have very similar logic. The base difference is the way of variable values division into groups – the *coarse classing* uses statistical methods to create groups in order to have a division including the most differentiate groups in the context of risk. Graphical results of *coarse classing* procedure are presented in the attachments to this paper.

During the estimation of the logistic regression model, popular algorithms for selecting the best model (selecting the best set of explanatory variables) is used, namely the backward, the forward and the stepwise methods. Additionally, the graphic visualization is used to present the values of the Bayesian information criterion (BIC) for various models (using different variables sets)⁵ that is presented in the figure 3.



Source: Own elaboration.

Figure 3. Values of BIC statistics for models estimated with using different variables sets⁶

Finally 6 models are estimated. Results of those models are presented in the table 5.

⁵ The initial set of variables is the set of all available 16 explanatory variables.

⁶ The BIC value on the graph means the approximate difference of this indicator between the BIC of the given model and the BIC of the model that includes only a constant. The bigger the difference, the better the model.

Table 5. Results of logistic models

Variable/model	model 1	model 2	model 3	model 4	model 5	model 6
<i>Intercept</i>	-1.6664*** (0.01)	-1.6664*** (0.01)	-1.6662*** (0.01)	-1.6664*** (0.01)	-1.6664*** (0.01)	-1.6639*** (0.0095)
<i>Acc_open_past_24mths_woe</i>	-0.6797*** (0.0673)	-0.6652*** (0.0654)	-0.6621*** (0.0654)	-0.6627*** (0.0654)	-0.7045*** (0.0643)	
<i>Dti_woe</i>	-0.1631** (0.0528)	-0.1628** (0.0528)	-0.1662** (0.0528)	-0.1845*** (0.0524)		
<i>Revol_util_woe</i>	-1.0875*** (0.0574)	-1.0863*** (0.0574)	-1.1303*** (0.0524)	-1.1336*** (0.0524)	-1.1728*** (0.0512)	
<i>Annual_inc_woe</i>	-1.5293*** (0.0586)	-1.5292*** (0.0586)	-1.5309*** (0.0586)	-1.5304*** (0.0586)	-1.585*** (0.0565)	
<i>Percent_bc_gt_75_woe</i>	-0.1495 (0.0801)	-0.149 (0.0801)				
<i>Loan_amnt_woe</i>	-0.6281*** (0.0705)	-0.6264*** (0.0705)	-0.6274*** (0.0704)	-0.6699*** (0.0687)	-0.6979*** (0.0683)	
<i>Num_tl_op_past_12m_woe</i>	-0.3802*** (0.0881)	-0.4324*** (0.0666)	-0.432*** (0.0666)	-0.4307*** (0.0666)	-0.4327*** (0.0666)	
<i>Mo_sin_rcnt_tl_woe</i>	-0.094 (0.1038)					
<i>Mths_since_recent_bc_woe</i>	-0.5631*** (0.0731)	-0.5737*** (0.0721)	-0.5785*** (0.0721)	-0.5767*** (0.0721)	-0.5581*** (0.0719)	
<i>Num_rev_tl_bal_gt_0_woe</i>	-0.4094*** (0.0734)	-0.4155*** (0.073)	-0.4354*** (0.0722)	-0.4348*** (0.0722)	-0.4673*** (0.0717)	
<i>Mort_acc_woe</i>	-0.7002*** (0.1102)	-0.7023*** (0.1101)	-0.7067*** (0.1101)	-0.6915*** (0.11)	-0.6804*** (0.1099)	
<i>Earliest_cr_line_n_woe</i>	-0.6441*** (0.1089)	-0.6466*** (0.1088)	-0.644*** (0.1088)	-0.6345*** (0.1088)	-0.6069*** (0.1085)	
<i>Term_woe</i>	-0.8989*** (0.0293)	-0.8996*** (0.0293)	-0.9003*** (0.0293)	-0.9196*** (0.0284)	-0.9212*** (0.0284)	
<i>Verification_status_woe</i>	-0.1554** (0.0584)	-0.1546** (0.0584)	-0.1541** (0.0584)			
<i>Inq_last_6mths_woe</i>	-0.8608*** (0.0538)	-0.8655*** (0.0535)	-0.8644*** (0.0535)	-0.8673*** (0.0535)	-0.871*** (0.0535)	
<i>Purpose_woe</i>	-0.861*** (0.0502)	-0.862*** (0.0502)	-0.862*** (0.0502)	-0.8693*** (0.0501)	-0.8681*** (0.0501)	
AIC	68626.8	68625.62	68627.11	68632.09	68642.52	73137.14
BIC	68785.44	68774.93	68767.08	68762.73	68763.83	73146.47

Information:

*** – p-value < 0.001

** – p-value < 0.01

* – p-value < 0.05

() – standard deviation

stepwise: model 3

forward: model 2

backward: model 2

Source: Own calculation.

The choice of the final model is based on the Akaike information criteria (AIC) and Schwartz information criteria (BIC). According to the convention, it is considered that the better model is this one that has the lowest value of the information criterion. AIC points that the best model is model no. 2, while BIC states that the best is model no. 4. However, the literature shows that the AIC criterion tends to consider a model with too many parameters as the best one. This is consistent with the analyzed case – model 2 contains more parameters in comparison to model 4. Therefore, model 4 is finally considered as the best logistic regression model.

The results of diagnostic tests for logistic regression model 4 are presented in the table 6.

Table 6. Results of diagnostic tests for model 4 (logistic regression)

Test	Statistic	p-value
Hosmer-Lemeshow (5 bins)	3.0610	0.3823
Hosmer-Lemeshow (10 bins)	10.8666	0.2094
Hosmer-Lemeshow (25 bins)	20.10210	0.0927
Osius-Rojek	-	0.7426
LR (likelihood ratio)	-	0.0000
ROC (model 4 vs. model 6)	-	0.0000
ROC (model 4 vs. model 1)	-	0.1982

Source: Own calculations.

The goodness of fit of the model to the data is checked using the Hosmer-Lemeshow test. The null hypothesis in this test is a statement of a good fit of the model to the data. In all the presented versions of this test there are no reason to reject the null hypothesis ($p\text{-value} > 0.05$) at the 5% significance level, therefore the model is well fitted to the data. However, this test has many disadvantages, including high sensitivity to the number of bins. That is why there are some critical opinions about this test (Allison 2013). For this reason, another test is carried out – the Osius-Rojek test that also verifies the goodness of fit of the model. According to the test, there are no reason to reject the null hypothesis about a good fit of the model to the data ($p\text{-value} > 0.05$) at a 5% level significance. The model specification can be considered as correct one. Additionally, the likelihood ratio (LR) test is carried out to verify significance of the all variables in the model (their total significance). According to the results of this test, the null hypothesis about the lack of significance of all variables is rejected ($p\text{-value} < 0.05$) at the 5% significance level, therefore it is assumed that the variables in the selected model are significant. Additionally, a formal check is done, using tests, to determine whether the ROC curves for models 4 and 6, as well as for models 4 and 1 are equally good. The null hypothesis is that ROC curves are equally good. The results of this test shows that ROC curves for models 4 and 6 are not equally good (and the information criteria indicates that model 4 is the better one). On the

other hand, for models 4 and 1, there are no reason to reject the null hypothesis about equally good ROC curve (p-value > 0.05) at the 5% significance level, therefore the ROC curves are equally good (and the BIC indicates that the better model is a model no. 4).

Results of tests that are described above met the expectations, therefore model 4 is considered as the best choice. In the next stage, the prediction is made – for all observations (in the training and the validation sets) the probability of default estimated by the model is assigned. Afterwards, the evaluation of model prediction is done by using different measures – results are shown in the table 7.

Table 7. Evaluation of model prediction – logistic regression

Measure	Training data	Validation data
AUC	0.6772	0.6708
Gini	0.3545	0.3416
KS	0.2580	0.2447

Source: Own calculations.

4.3. Neural networks model

In the practical approach to the building of neural networks, in the literature it can be found that the best results can be obtained when all variables (input and output) have similar ranges of values – otherwise, weights assigned to variables with higher values may dominate the network training process (Szeliga, 2017). Therefore, continuous explanatory variables have been standardized according to the formula (Walesiak, 2011):

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j},$$

where:

z_{ij} – normalized value of the j -th variable for the i -th record,

x_{ij} – value of j -th variable for i -th record,

$\bar{x}_j$ – average of j -th variable,

s_j – standard deviation of j -th variable.

Standardization of continuous variables lets to have a fixed range of variables' values (from 0 to 1). In addition, basing on categorical variables many binary variables are created – separate variables are generated for each variable's category. The original values of the variables are not taken into account. It lets to have all variables' values in the range from 0 to 1 in the neural network modelling process. The dataset that is used to build the neural network model is analogical to this one that is used to estimate the logistic regression model in the context of variables set. However, after the transformations that are described above it has 35

characteristics (number of variables is bigger because of variables transformation – from many categories to binary form) – 1 target and 34 explanatory.

According to information from the section no. 2, in the context of neural network the most popular type is considered – multilayer perceptron (MLP). An important issue in constructing this type of neural network is to set the number of layers and the number of neurons in each layer. It is possible to find the information that an MLP network with two hidden layers can model almost any problem which does not mean that a neural network with more layers would not solve this problem in an easier or more convenient way – it is obvious that the more complex neural work (in terms of its structure) is able to return better results. In practice it is enough to use only one hidden layer for the majority of considered problems. In special cases, there is a need to include two hidden layers. The use of three hidden layers is extremely rare and in practice it is almost never used (StatSoft, 2006). It justifies the wide use of a simple neural network called *vanilla* – a network with one hidden layer, also known as a single-layer perceptron (Hastie *et al.*, 2008). In this paper a neural network with just one hidden layer is used. The problem that appears in the next step is the choice of the number of neurons in the hidden layer. There are many suggestions that can be helpful in choosing the right number of hidden neurons, including one thumb rule assumes that the number of hidden neurons should be between the number of neurons in the input layer and the number of neurons in the output layer (Panchal F. and Panchal M., 2014). In addition, sometimes the trial and error method is also applied, taking into account the *forward* algorithm, according to which it starts by considering a small number of hidden neurons (e.g. 2), then trains the network and checks its quality. Thereafter iteratively next neuron is added and the same set of actions is done until the neural network does not improves the results (Panchal F. and Panchal M., 2014). In this work two approaches are combined – the starting point is 2 hidden neurons. Then the neural network is trained, predictions are done and quality of model is checked. Iteratively all the procedure of adding another neuron (and other activities) is repeated until the variant with 35 neurons was checked (using the cited thumb rules). Due to the fact that classification is the problem solved by the neural network model in the analyzed case, the AUC measure calculated on the training set is used as the criterion for selecting the best network structure. Table 8 presents AUC values for individual network structures (with different numbers of neurons in the hidden layer) on the training sample.

Table 8. AUC values depending on number of neural neurons in hidden layer (training set)

Number of neurons	AUC	Number of neurons	AUC	Number of neurons	AUC
2	0.66533	13	0.67141	24	0.67137
3	0.66632	14	0.67092	25	0.67159
4	0.66621	15	0.67193	26	0.67201
5	0.66891	16	0.67289	27	0.67132
6	0.66856	17	0.6732	28	0.67297
7	0.66905	18	0.66898	29	0.67315
8	0.67048	19	0.67261	30	0.67254
9	0.67051	20	0.67251	31	0.67335
10	0.66943	21	0.67239	32	0.67364
11	0.67136	22	0.67259	33	0.67330
12	0.67061	23	0.67219		

Source: Own calculations.

Results presented in table 9 shows that the best neural network (in terms of AUC measure calculated on training set) has 32 neurons in the hidden layer and this model is chosen as the best one. Then a prediction is done and measures of model classification are calculated (for the training and validation set) – results are presented in table 9.

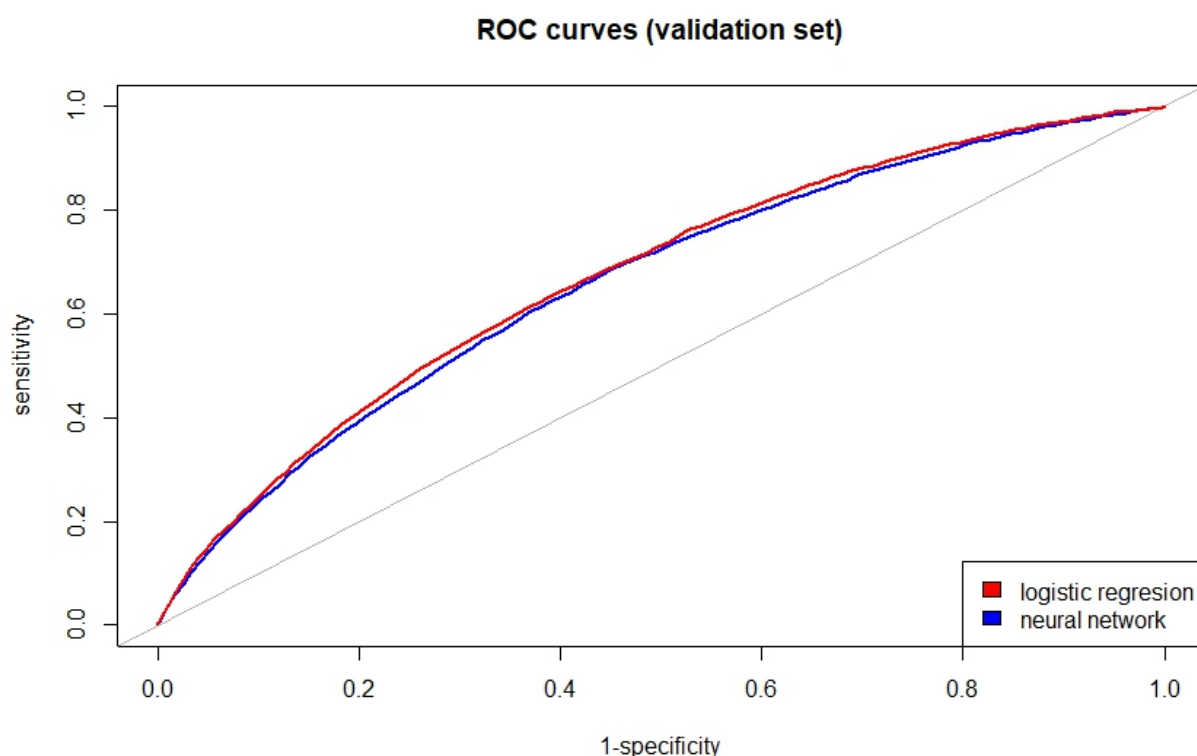
Table 9. Evaluation of model prediction – neural network

Measure	Training set	Validation set
AUC	0.6736	0.6601
Gini	0.3473	0.3202
KS	0.2557	0.2359

Source: Own calculations.

4.4. Models comparison

Logistic regression model and neural network model are estimated. It is very important to determine which model is better in terms of prediction. At the beginning, the classification capabilities of models are compared graphically by using ROC curves for both models (based on the validation set) that are presented in the figure 4.



Source: Own elaboration.

Figure 4. ROC curves for final models (validation set)

In the table 10 there are measures to evaluate models classification (based on validation set).

Table 10. Evaluation of models classification (validation set; in %) ⁷

Measure/model	Logistic regression			Neural network		
	$Z^*=10$	$Z^*=25$	$Z^*=50$	$Z^*=10$	$Z^*=25$	$Z^*=50$
AUC	67.72			66.01		
Gini	35.45			32.02		
KS	25.80			23.59		
Accuracy	38.98	78.71	84.07	34.01	76.14	84.11
Error ratio	61.02	21.29	15.93	65.99	23.86	15.89
PPV	19.2	31.07	48.91	18.28	28.62	53.05
NNV	93.05	86.58	84.16	92.9	86.97	84.25

Source: Own elaboration.

Based on the ROC curves graph, it can be noticed that probably logistic regression model does better classifies borrowers into two groups (clients repaying their loans on time and borrowers having problems with it) because the logistic regression's ROC curve line is more shifted towards point (1, 1) in comparison to the neural network's ROC curve. Similar conclusions are based on numerical analysis. According to the measures presented in the table 10, it can also be thought that the logistic regression model is better than the neural network

⁷ Some measures depending on cutoff point (marked as Z^*).

model – the AUC, Gini and KS statistics are higher for regression model than for neural network model, what shows better prediction power of logistic regression model.

What is more, the formal test is carried out to verify that the ROC curves for the compared models are equally good (the null hypothesis assumes that ROC curves are equally good). Based on the test results that are obtained, the null hypothesis is rejected (at a 5% significance level). Therefore, there are statistically significant differences in the classification effectiveness of the analyzed models.

Analyzing measures from table 11 allows to notice that choice of cutoff point (defining boundary between bad and good cases) is extremely important. When cut-off point increases, the accuracy and precision of the positive prediction also increases and, consequently, the classification error and precision of the negative prediction decreases (it is worth to notice that regardless of the cut-off point considered, the effectiveness of logistic regression model is better than neural network model). This is a key element in a business environment, especially in financial institutions, because it relates to controlling the sales process (loans granted), as well as the quality of the loan portfolio (having a specified percentage of clients who do not pay their liabilities accordingly with the contract). That is why an econometric model that have very good prediction power is a very crucial and practical business tool.

5. Conclusions

Financial institutions engaged in activities based on lending money play a key role in everyone's life, because they improve the quality of life. In addition, the condition of financial institutions influences the functioning of the finances of the entire country. Recently, an alternative way of borrowing money - social loans, is gaining quite significant importance. An important element of all financial institutions involved in credit activity is strict control of credit risk (its minimization), which translates into the financial condition of the institution.

In the era of the growing importance of data and advanced data analysis, companies have a number of methods at their disposal to optimize the management of various processes. Various types of econometric models are used to properly measure and manage credit risk. Many studies compare the effectiveness of various techniques to enable the correct classification of good customers (loan repayments on time) and bad customers (loan repayments on time). The choice of the best classifier plays a key role in controlling the percentage of non-performing loans in the portfolio of a given financial institution, which translates into its functioning.

Our paper reviews literature in the context of comparison of classifier results. The two most frequently compared techniques are the logistic regression model and the neural network model. In our research we analyze these two techniques on real social lending data. The aim of the study was to verify the hypothesis that neural nets model is a better classifier compared to logistic regression. Our hypothesis was rejected. We found that logistic regression performs better in classifying customers as good or bad. Our results on data from the popular social lending platform Lending Club indicate the logistic regression outperforms neural nets. Similar conclusions were stated by Bekhet and Eletter (2014).

The reason for our result may be the use of WOE transformation in the construction of a logistic regression model, which certainly improved its results, and was not always used by other researchers in the context of comparing classifiers. In addition, the poor efficiency of neural networks may have resulted from the implementation of a fairly simple network structure, which, although widely used, could prove less effective than more complex network structures, including more hidden layers.

References

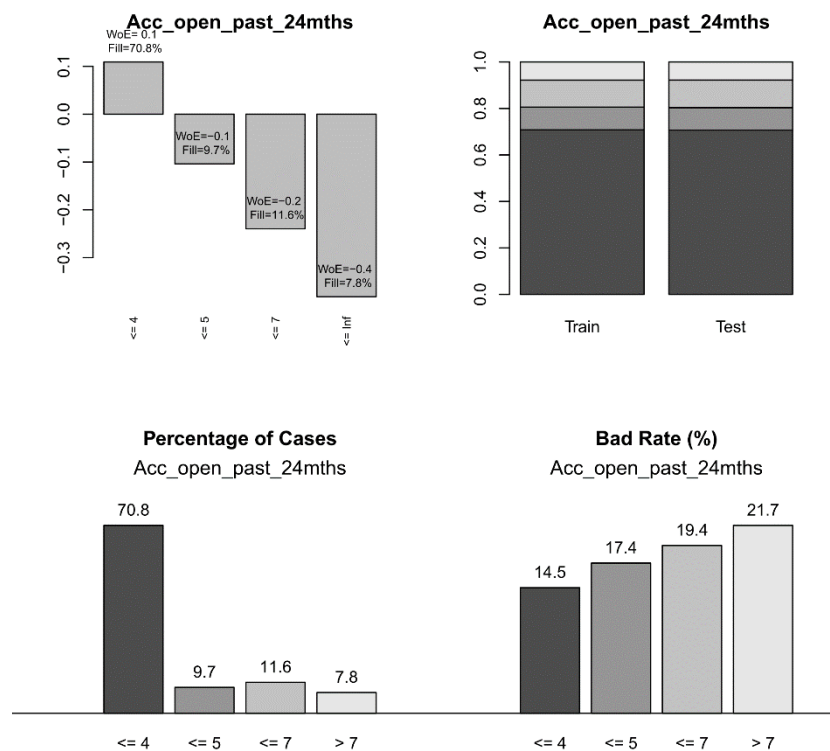
- Abdou H., Pointon J., El-Masry A., 2008. Neural nets versus conventional techniques in credit scoring in Egyptian banking. *Expert Systems with Applications*, 35, 1275-1292.
- Abdou H., Dongmo Tsafack M., Ntim C., Baker R., 2016. Predicting creditworthiness in retail banking with limited scoring data. *Knowledge-Based Systems*, 103, 89-103.
- Akkoç S., 2012. An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data. *European Journal of Operational Research*, 222, 168-178.
- Allison P., 2013. Why I Don't Trust the Hosmer-Lemeshow Test for Logistic Regression. *Statistical Horizons*, <https://statisticalhorizons.com/hosmer-lemeshow> [accessed 27.08.2018].
- Bekhet H., Eletter S., 2014. Credit risk assessment model for Jordanian commercial banks: Neural scoring approach. *Review of Development Finance*, 4, 20-28.
- Bogges W. P., 1967. Screen-test your credit risks. *Harvard Business Review* 45, 113-122.
- Bogus A., 2002. Kredyty w teorii racjonalnego wyboru konsumpcji. *Acta Universitatis Lodzensis. Folia Oeconomica*, 161, 257-265.
- Chang S., Kim S. D., Kondo G., 2015. Predicting Default Risk of Lending Club Loans. *Machine Learning*, CS229, 1-5.
- Chrupczalski S., 2010. *Asymetria informacji*. Narodowy Bank Polski – Portal Edukacji Ekonomicznej, <https://www.nbportal.pl/wiedza/artykuly/na-poczatek/asymetria-informacji> [accessed 27.08.2018].
- Cramer J., 2004. The early origins of the logit model. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 35, 613-626.
- Dziuba D., 2017. Rozwój systemów pożyczek społecznościowych determinowany potrzebami użytkownika IT (konsumenta). *Problemy Zarządzania*, 4 (71), 116-126.
- Emekter R., Tu Y., Jirasakuldech B., Lu M., 2015. Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. *Applied Economics*, 47 (1), 54-70.
- Folwarski M., 2016. Wpływ kredytów bankowych na wzrost gospodarczy w Polsce. *Zarządzanie Finansami i Rachunkowość*, 4 (3), 5-14.
- Fritzpatrick T., Mues C., 2016. An empirical comparison of classification algorithms for mortgage default prediction: evidence from a distressed mortgage market. *European Journal of Operational Research*, 249, 427-439.
- Gruszczyński M., Kuszewski T., Podgórska M., 2009. *Ekonometria i badania operacyjne*. PWN.

- Gwizdała J., 2011. *Ryzyko kredytowe w działalności banku komercyjnego*. Wydawnictwo Uniwersytetu Gdańskiego.
- Hand D., Till R., 2001. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, 45, 171-186.
- Hastie T., Tibshirani R., Friedman J., 2008. *The Elements of Statistical Learning – Data Mining, Inference, and Prediction*. Springer, second edition.
- Haykin S., 1999. *Neural networks – A comprehensive foundation*. Pearson Education, second edition.
- Imtiaz S., Brimicombe A., 2017. A Better Comparison Summary of Credit Scoring Classification. *International Journal of Advanced Computer Science and Applications*, 8 (7), 1-4.
- Ince H., Aktan B., 2009. A comparison of data mining techniques for credit scoring in banking: A managerial perspective. *Journal of Business Economics and Management*, 10 (3), 233-240.
- Kočenda E., Vojtek M., 2011. Default Predictors in Retail Credit Scoring: Evidence from Czech Banking Data. *Emerging Markets Finance and Trade*, 47 (6), 80-98.
- Kopczewska K., 2011. *Ekonometria i statystyka przestrzenna z wykorzystaniem programu RCRAN*. CeDeWu, first edition.
- Korzeniowska A., 2012. Rozwój pożyczek społecznościowych w Polsce. *Annales Universitatis Mariae Curie-Skłodowska. Sectio H, Oeconomia*, 4, 427-436.
- Lahsasna A., Ainon R.N., The Y.W., 2010. Credit Scoring Models Using Soft Computing Methods: A Survey. *The International Arab Journal of Information Technology*, 7, 115-123.
- Malekipirbazari M., Aksakalli V., 2015. Risk assessment in social lending via random forests. *Expert Systems with Applications*, 42, 4621-4631.
- Matuszyk A., 2018. *Credit Scoring*. CeDeWu, second edition.
- Mitrega-Niestrój K., 2012. Bankowość p2p jako alternatywa dla tradycyjnych banków detalicznych. *Annales Universitatis Mariae Curie-Skłodowska. Sectio H, Oeconomia*, 4, 601-608.
- Mycielski J., 2010. *Ekonometria*. Sowa – druk na życzenie, third edition.
- Nehrebecka N., Dzik A., 2012. Konstrukcja miernika szans na bankructwo firmy. *Materiały i studia*. Narodowy Bank Polski, 280.

- Panchal F., Panchal M., 2014. Review on Methods of Selecting Number of Hidden Nodes in Artificial Neural Network. *International Journal of Computer Science and Mobile Computing*, 3 (11), 455-464.
- Przanowski K., 2015. *Credit scoring. Studia przypadków procesów biznesowych*. Szkoła Główna Handlowa w Warszawie – Oficyna Wydawnicza.
- Przenajkowska K., 2008. Istota i znaczenie zarządzania ryzykiem kredytowym w obliczu załamania na amerykańskim rynku kredytów hipotecznych. *Equilibrium*, 1-2 (1), 113-126.
- Řezáč M., Řezáč F., 2011. How to Measure the Quality of Credit Scoring Models. *Czech Journal of Economics and Finance*, 61 (5), 486-507.
- Sawicka A., Tymoczko I., 2014. Dlaczego polskie przedsiębiorstwa nie korzystają z kredytu. Zjawisko zniechęconego kredyto(nie)biorcy. *Materiały i Studia*, 310.
- Serrano-Cinca C., Gutiérrez-Nieto B., López-Palacios L., 2015. Determinants of Default in P2P Lending. *PLoS ONE*, 10 (10), 1-22.
- Siddiqi N., 2006. *Credit risk scorecards: developing and implementing intelligent credit scoring*. Wiley.
- Słownik języka polskiego PWN, <https://sjp.pwn.pl/sjp/ryzyko;2518509.html> [accessed 27.08.2018].
- StatSoft, 2006. Sieci neuronowe. *Elektroniczny Podręcznik Statystyki PL*, <https://www.statsoft.pl/textbook/stathome.html> [accessed 27.08.2018].
- Szalaniec M., 2004. *Historia życia na kredyt*. Narodowy Bank Polski – Portal Edukacji Ekonomicznej, https://www.nbportal.pl/wiedza/artykuly/historia-mysli-ekonomicznej/historia_zycia_na_kredyt [accessed 27.08.2018].
- Szeliga M., 2017. *Data science i uczenie maszynowe*. PWN, first edition.
- Tadeusiewicz R., 2015. Szalaniec M. *Leksykon sieci neuronowych*. Projekt Nauka, first edition.
- Tsai K., Ramiah S., Singh S., 2014. *Peer Lending Risk Predictor*, CS229, 1-5.
- Tsai MC., Lin SP., Cheng CC., Lin YP., 2009. The consumer loan default prediction model – An application of DEA-DA and neural network. *Expert Systems with Applications*, 36, 11682-11690.
- Walesiak M., 2011. *Uogólniona miara odległości GDM w statystycznej analizie wielowymiarowej z wykorzystaniem programu R*. Wydawnictwo Uniwersytetu Ekonomicznego.
- Wang G., Hao J., Ma J., Jiang H., 2011. A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, 28, 223-230.

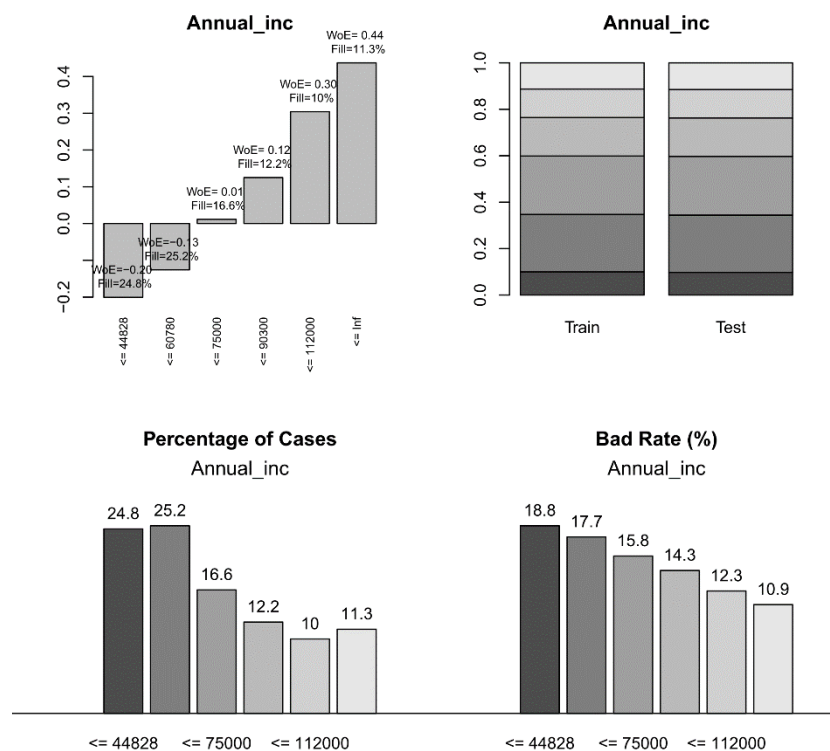
- Yeh IC., Lien C., 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36, 2473-2480.

Appendix – Coarse classing procedure results for particular variables



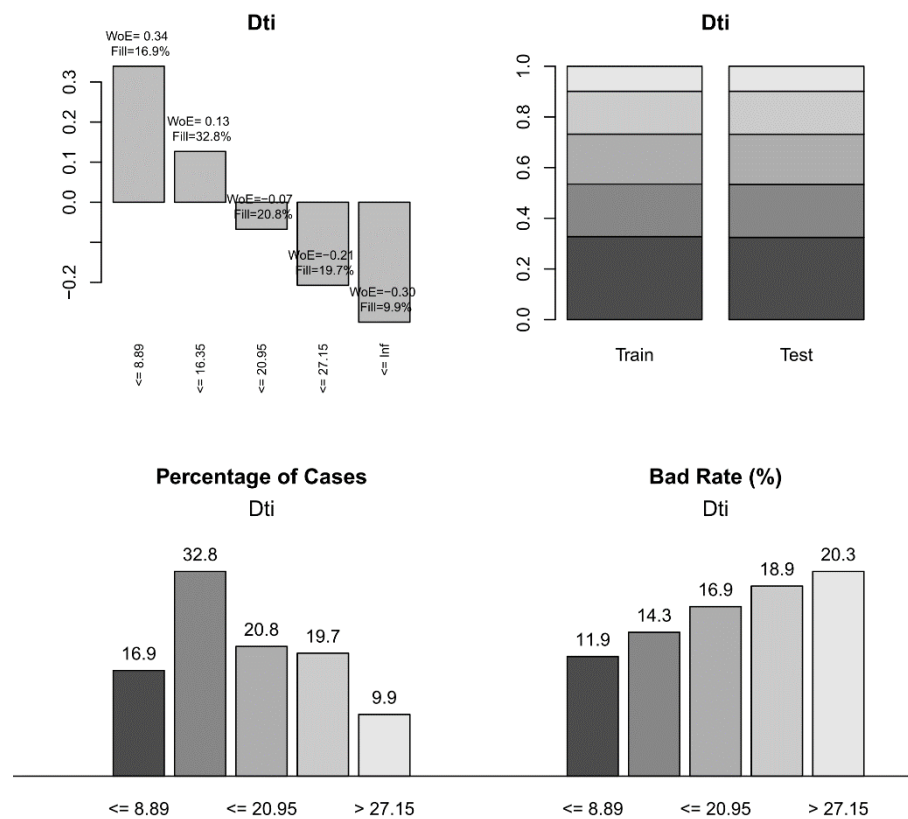
Source: Own elaboration.

Figure 1. Variable Acc_open_past_24mths – coarse classing



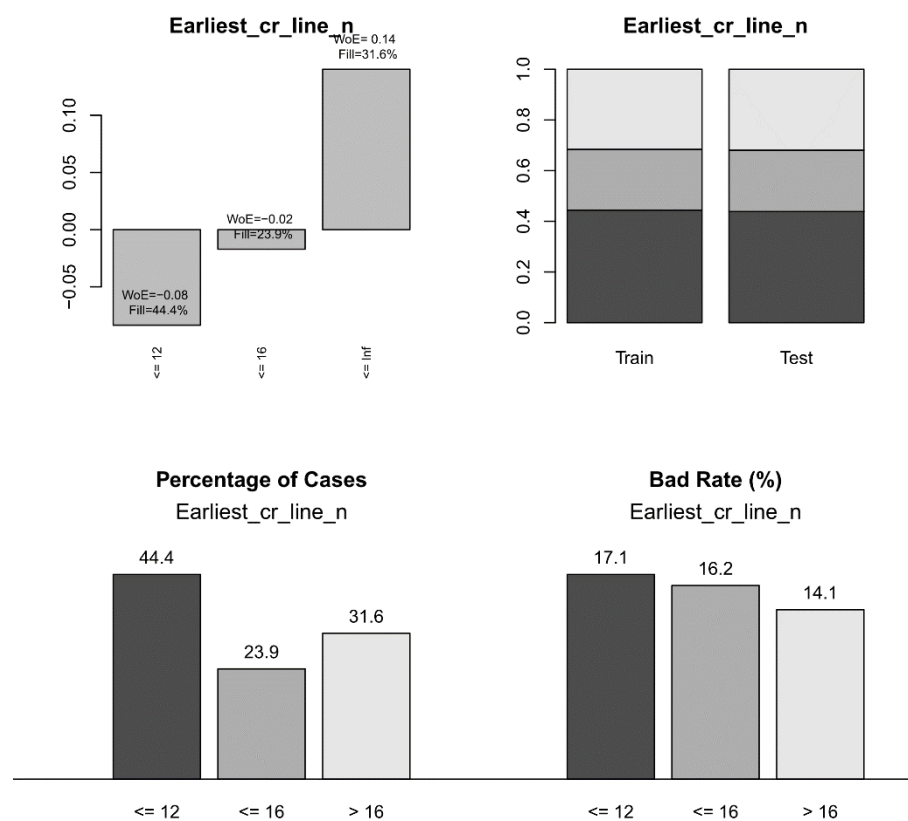
Source: Own elaboration.

Figure 2. Variable Annual_inc – coarse classing



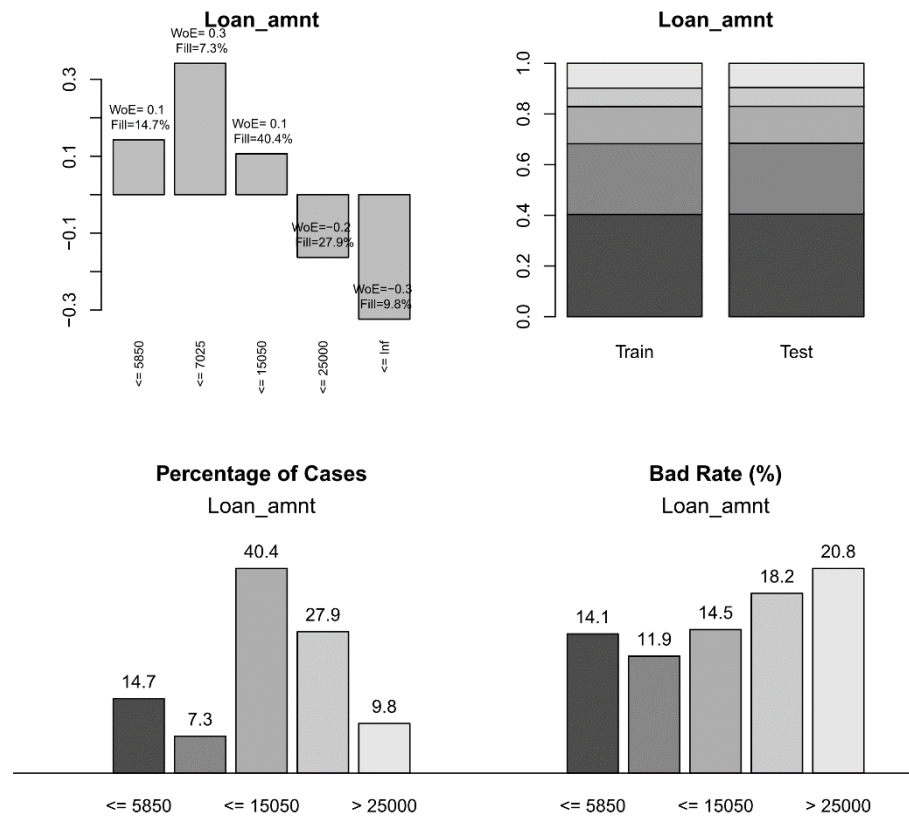
Source: Own elaboration.

Figure 3. Variable Dti – coarse classing



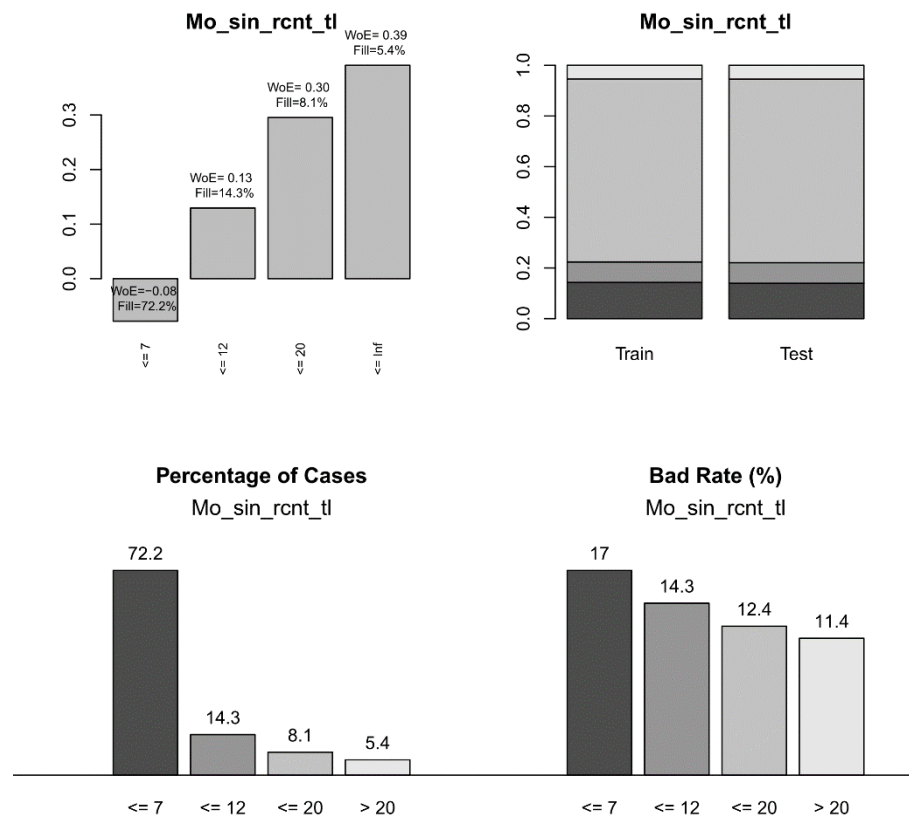
Source: Own elaboration.

Figure 4. Variable Earliest_cr_line_n – coarse classing



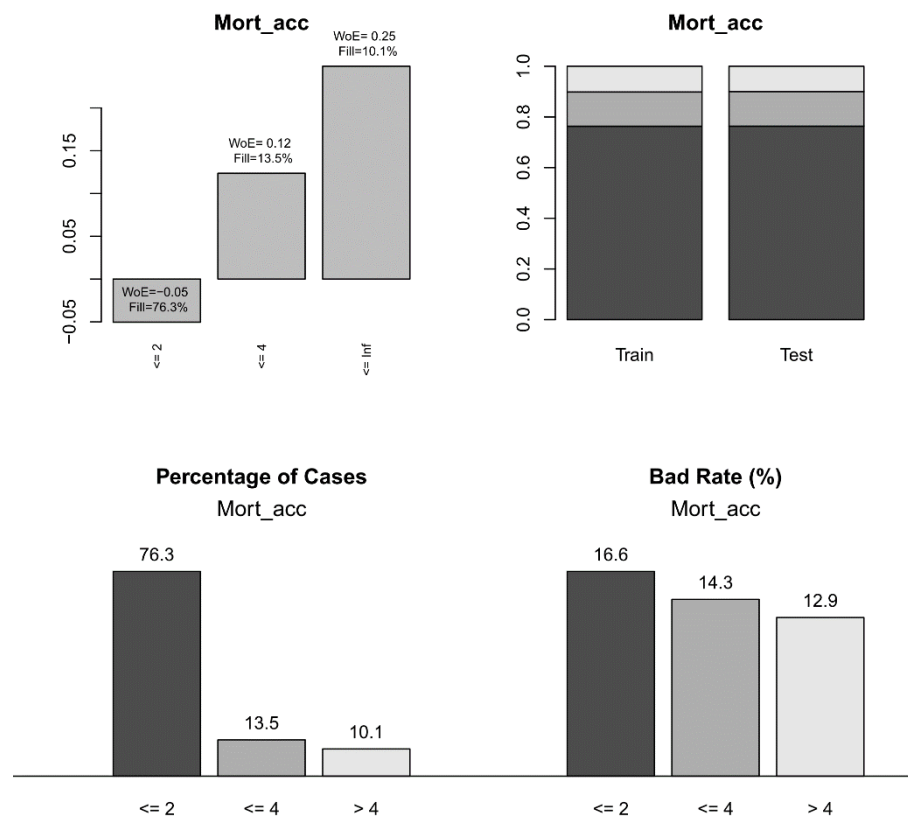
Source: Own elaboration.

Figure 5. Variable Loan_amnt – coarse classing



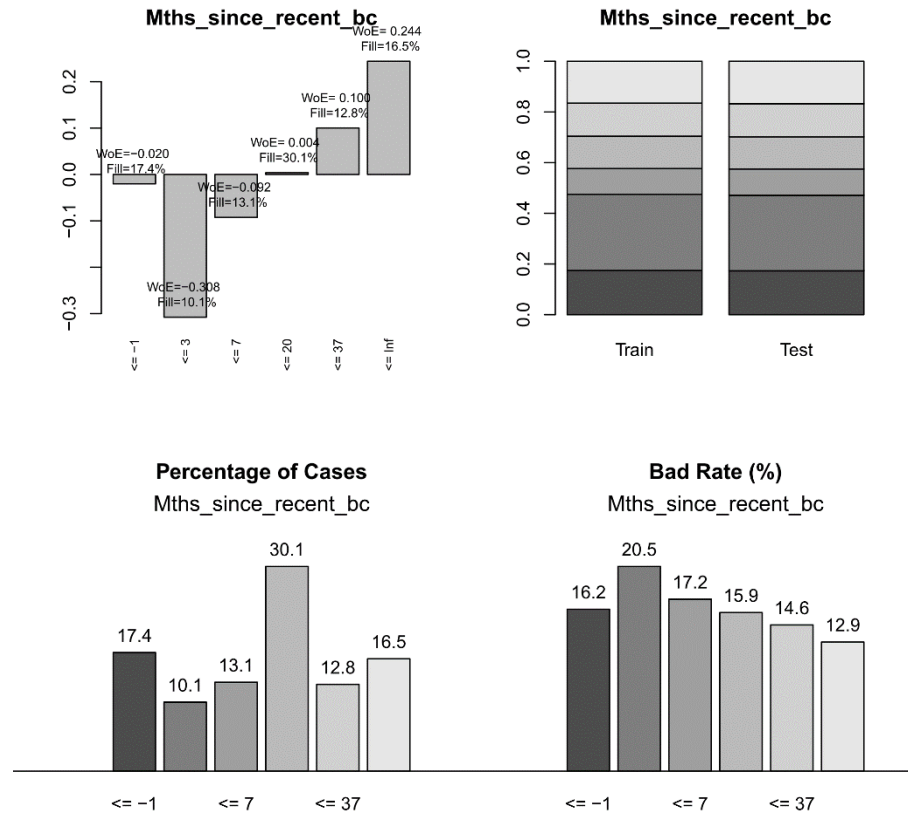
Source: Own elaboration.

Figure 6. Variable Mo_sin_rcnt_tl – coarse classing



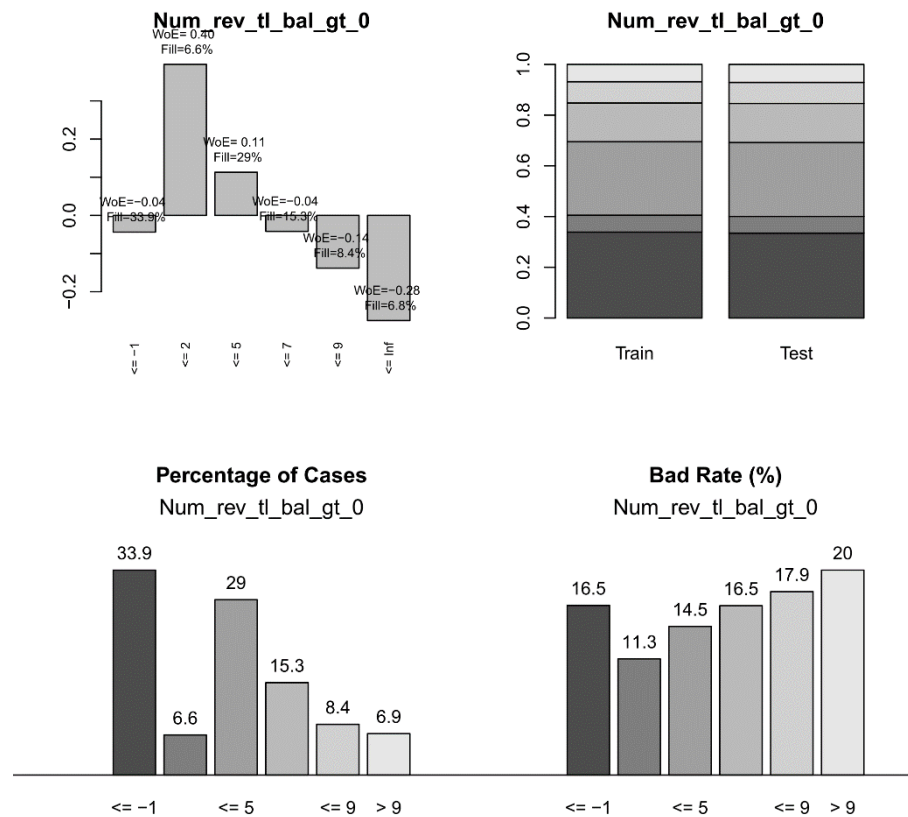
Source: Own elaboration.

Figure 7. Variable Mort_acc – coarse classing



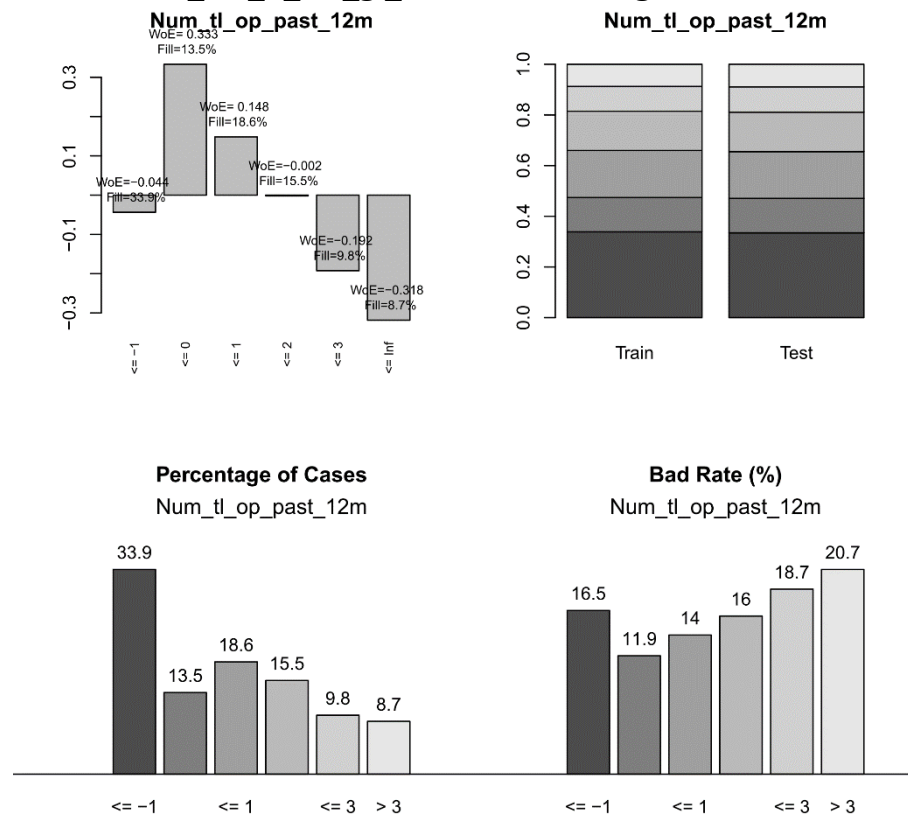
Source: Own elaboration.

Figure 8. Variable Mths_since_recent_bc – coarse classing



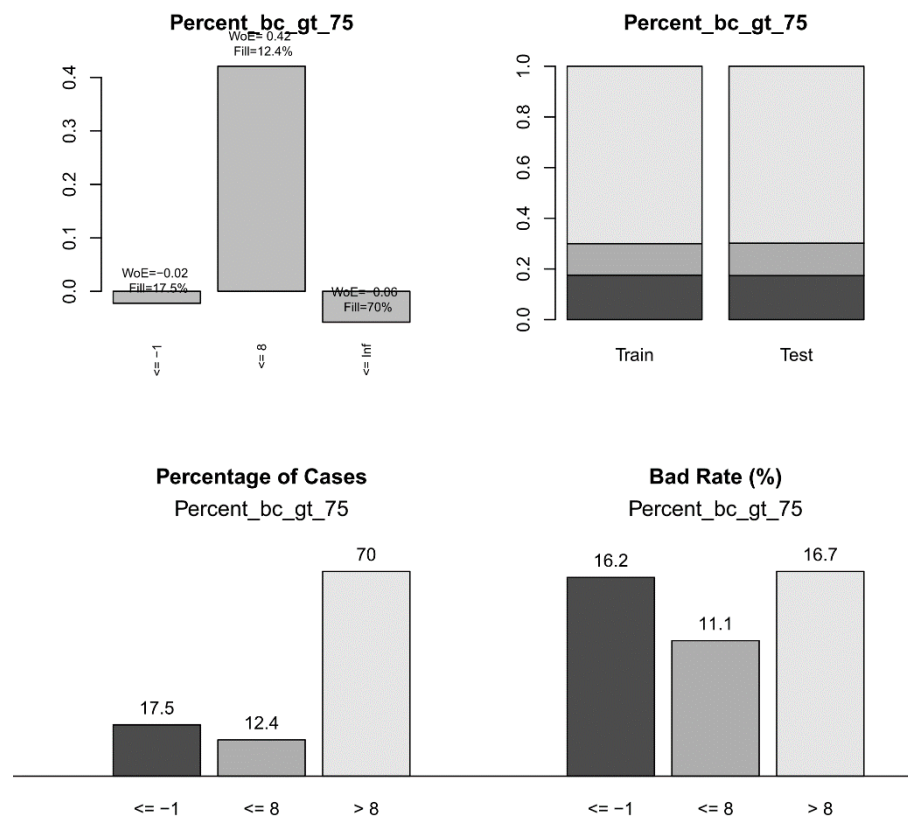
Source: Own elaboration.

Figure 9. Variable Num_rev_tl_bal_gt_0 – coarse classing



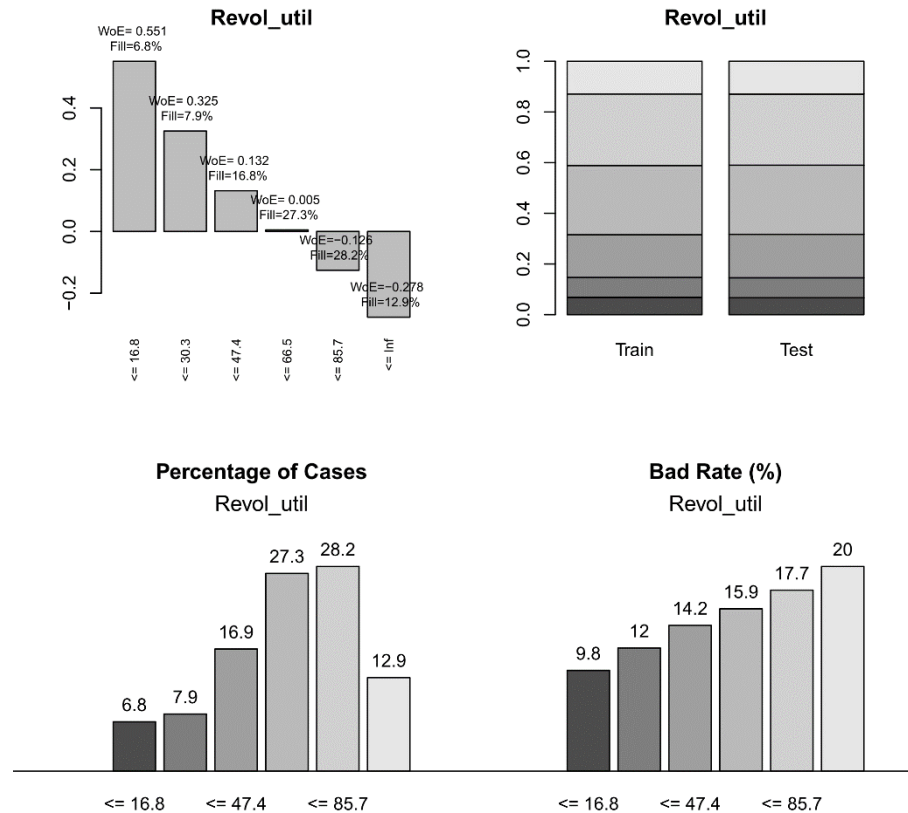
Source: Own elaboration.

Figure 10. Variable Num_tl_op_past_12m – coarse classing



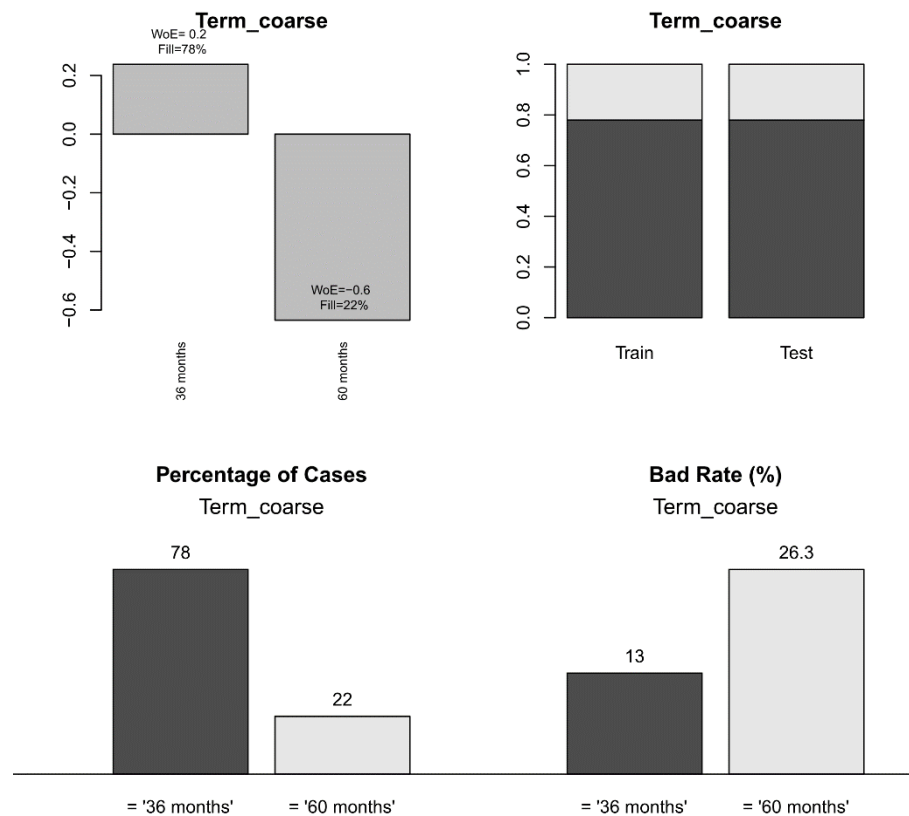
Source: Own elaboration.

Figure 11. Variable Percent_bc_gt_75 – coarse classing



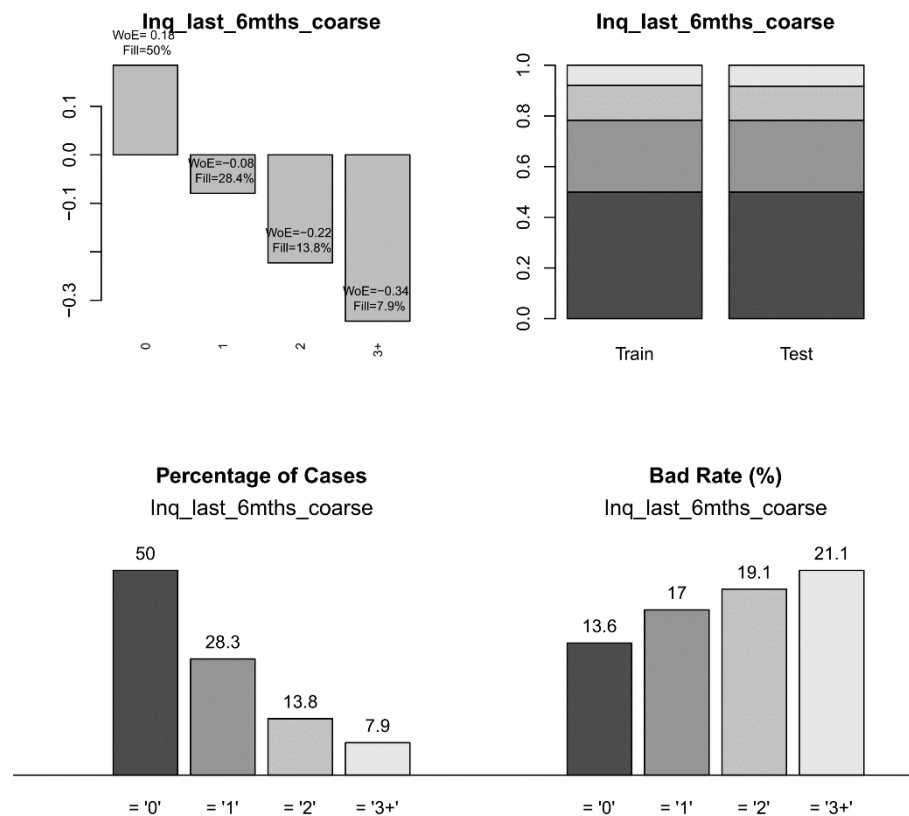
Source: Own elaboration.

Figure 12. Variable Revol_util – coarse classing



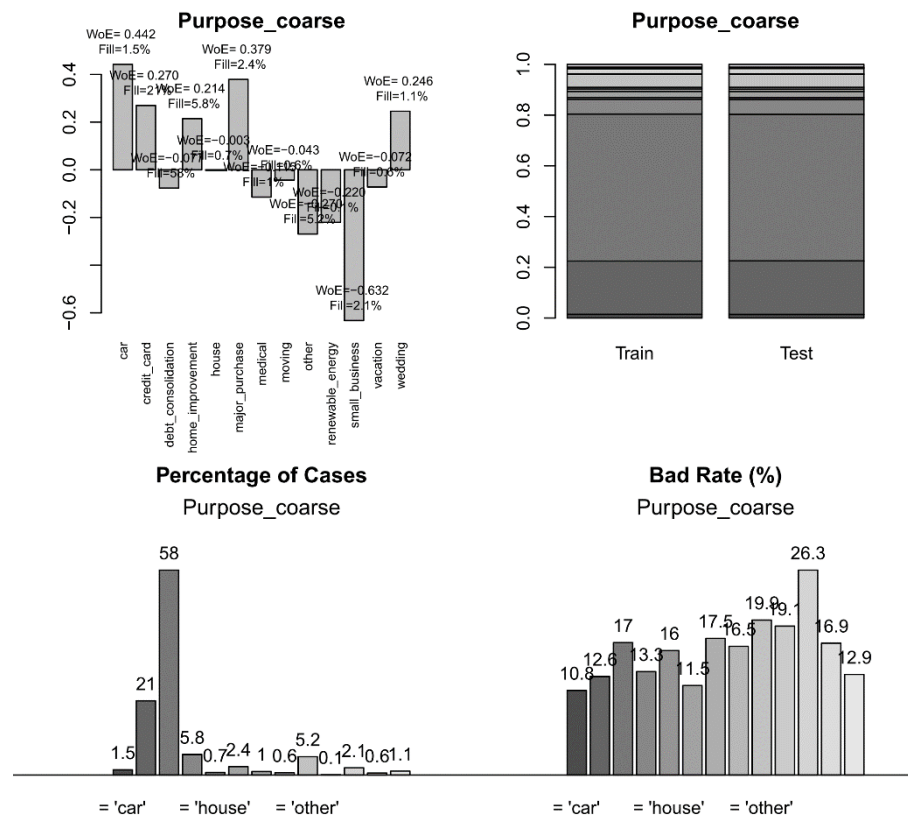
Source: Own elaboration.

Figure 13. Variable Term – coarse classing



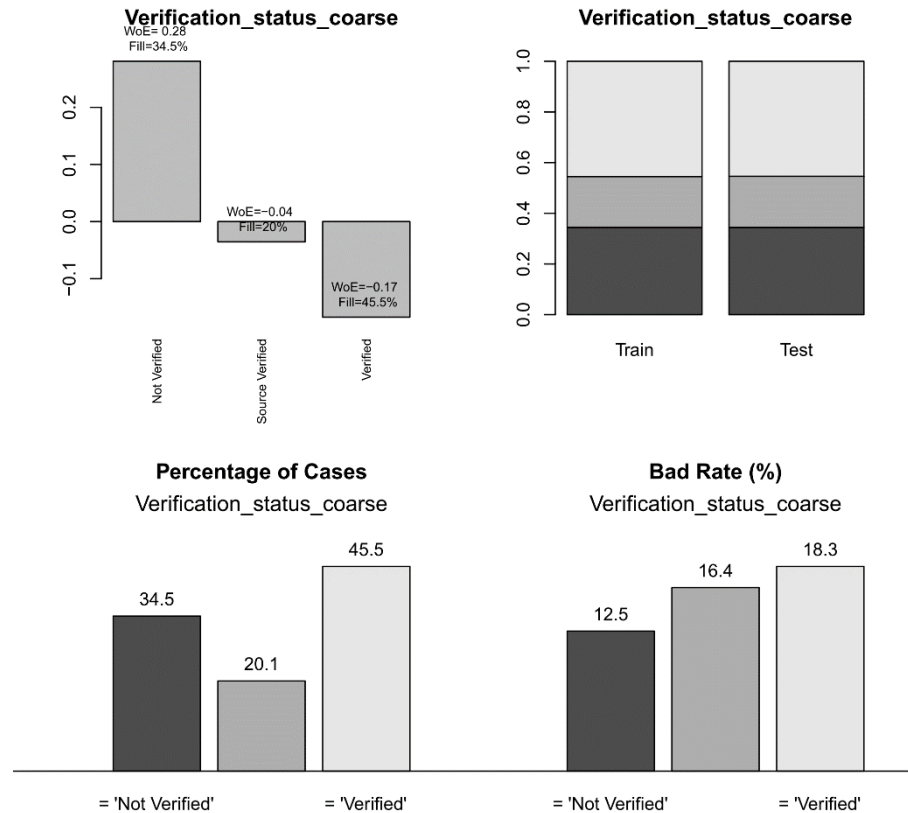
Source: Own elaboration.

Figure 14. Variable Inq_last_6mths – coarse classing



Source: Own elaboration.

Figure 15. Variable Purpose – coarse classing



Source: Own elaboration.

Figure 16. Variable Verification_status – coarse classing



UNIVERSITY OF WARSAW

FACULTY OF ECONOMIC SCIENCES

44/50 DŁUGA ST.

00-241 WARSAW

WWW.WNE.UW.EDU.PL