Preliminary communication
UDC: 336.77:519.8
https://doi.org/10.18045/zbefri.2021.1.163

# An implementation of ensemble methods, logistic regression, and neural network for default prediction in Peer-to-Peer lending[*]

*Aneta Dzik-Walczak*[1], *Mateusz Heba*[2]

***Abstract***

*Credit scoring has become an important issue because competition among financial institutions is intense and even a small improvement in predictive accuracy can result in significant savings. Financial institutions are looking for optimal strategies using credit scoring models. Therefore, credit scoring tools are extensively studied. As a result, various parametric statistical methods, non-parametric statistical tools and soft computing approaches have been developed to improve the accuracy of credit scoring models. In this paper, different approaches are used to classify customers into those who repay the loan and those who default on a loan. The purpose of this study is to investigate the performance of two credit scoring techniques, the logistic regression model estimated on categorized variables modified with the use of WOE (Weight of Evidence) transformation, and neural networks. We also combine multiple classifiers and test whether ensemble learning has better performance. To evaluate the feasibility and effectiveness of these methods, the analysis is performed on Lending Club data. In addition, we investigate Peer-to-peer lending, also called social lending. From the results, it can be concluded that the logistic regression model can provide better performance than neural networks. The proposed ensemble model (a combination of logistic regression and neural network by averaging the probabilities obtained from both models) has higher AUC, Gini coefficient and Kolmogorov-Smirnov statistics compared to other models. Therefore, we can conclude that the ensemble model*

[*] Received: 08-04-2020; accepted: 15-03-2021

[1] *Assistant Professor, University of Warsaw – Faculty of Economic Sciences, Długa 44/50, 00-241 Warsaw, Poland. Scientific affiliation: econometric methods and models, corporate finance. Phone: +48 22 55 49 111. Fax: 22 831 28 46. E-mail: adzik@wne.uw.edu.pl. Personal website: http://www.wne.uw.edu.pl/index.php/pl/profile/view/155/.*

[2] *PhD Student, University of Warsaw – Faculty of Economic Sciences, Długa 44/50, 00-241 Warsaw, Poland. Scientific affiliation: econometric methods and models, financial risk, credit risk, corporate finance. Phone: +48 22 55 49 111. Fax: 22 831 28 46. E-mail: mheba@wne.uw.edu.pl. Personal website: https://www.wne.uw.edu.pl/index.php/en/profile/ view/336/.*

*allows to successfully reduce the potential risks of losses due to misclassification costs.*

***Key words:*** *credit scoring, ensemble methods, logistic regression, neural nets, peer-to-peer lending*

***JEL classification:*** *G21, G32*

## 1. Introduction

Credit risk is an inseparable aspect of lending. An important issue for the financial institution is to achieve the lowest possible percentage of non-performing loans by minimizing the information asymmetry between the lender and the borrower. Credit risk decisions are a critical factor in the success of financial institutions due to the very high cost of bad decisions (Lahsasna et al., 2010). Credit risk assessment is an essential element of credit risk management and forms the basis for credit decisions (Wu et al., 2010). Due to the importance of credit risk, several tools have been proposed to improve risk scoring and increase predictive accuracy. Credit scoring aims to classify customers as good customers, i.e., customers who repay the loan, and bad customers, i.e., customers who default on a loan. Various 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 assess credit risk.

Peer-to-peer (P2P) lending has developed in recent years. In P2P lending, also called social lending, individuals lend money to other individuals. There is no financial institution involved in this process. In P2P lending, individual investors bear the credit risk, not financial institutions. An electronic platform mediates between borrowers and lenders and charges a fee for the service. Companies broker loans between individuals.

The research question of this paper aims to compare the following credit score classification methods: logistic regression model and neural networks, and a combination of these models using ensemble methods. In recent years, neural networks are perceived as one of the best statistical techniques for building scoring models. Moreover, the use of ensemble methods very often improves the performance of the classifier compared to single methods. We test the hypothesis that the neural network model is a better classifier compared to logistic regression, but using ensemble methods to combine these two single classifiers allows us to increase the final predictive power of the model. We use data from Lending Club, the largest US P2P lending company. The analyzed sample contains 119,160 loans within the period 2011 – 2013.

The rest of the paper is organized as follows. In Section 2, we present the literature review. In Section 3, we describe the details of the methodology. In

Section 4, we present data and analysis. Section 5 presents the empirical results and discussion, while in Section 6, we present the conclusions based on the results of our research.

## 2. Literature review

Credit scoring is a financial tool used in the process of risk assessment, namely to manage and diversify risk in investment portfolios. One of the main objectives of this tool is to assess the risk of loans. Credit scoring models allow the classification of loan customers into a good or bad category in terms of their characteristics and are widely used by financial institutions. They are used to score new consumer loans as well as to analyze existing loans. Credit scoring models make it possible to reduce the cost of credit analysis, enable faster credit decisions, and reduce potential risks (Lee et al., 2002; West, 2000). Scoring tasks are then associated with classification analysis. Since improving such classification can lead to significant savings, the accuracy of different techniques should be analyzed and compared.

In order to increase the accuracy of the credit scoring model, numerous methods have been developed. Both parametric statistical techniques (e.g. discriminant analysis, logistic regression) and non-parametric statistical techniques (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 the literature review, it can be concluded that there is no general best method used in building credit scoring models. Authors state that many factors such as data structure, explanatory variables, comparison criterion are relevant in selecting the best technique of classification (e.g. Hand and Henley, 1997).

Abdou et al. (2008) find that the criterion of lowest misclassification cost leads network search to select a multilayer feed-forward network with five nodes. However, probabilistic neural networks provide the highest average correct classification rate. Tsai et al. (2009) show that DEA -DA (Data Envelopment Analysis-Discriminant Analysis) and neural networks have better predictive ability than probit analysis and logistic regression. Neural networks were also selected as the optimal predictive model by Yeh and Lien (2009). Kočenda and Vojtek (2011) found that both credit risk models based on logistic regression and regression trees are comparably efficient. Wang et al. (2011) studied logistic regression analysis, decision tree, artificial neural network and support vector machine. The bagging decision tree achieved 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 that performs better than linear discriminant analysis, logistic regression and artificial neural network in terms of average correct classification rate and estimated misclassification cost. Bekhet and Eletter (2014) state that logistic regression performs better than the radial basis function model in terms of overall accuracy rate. However, the radial basis function is superior in identifying customers who may drop out. Tsai et al. (2014) attempted to reduce the risk of loan defaults and outperform the return of Lending Club at the same level of risk. They used four algorithms: Naive Bayes, Random Forest, Support Vector Machines, and modified logistic regression (penalty if the classifier misclassifies a defaulted loan as a good loan). The highest precision was obtained for modified logistic regression on two-dimensional principal component analysis data. Chang et al (2015) compared the 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, G-mean, completion of performance metrics, accuracy and precision. They found that Naive Bayes with Gaussian performs the best in standard prediction (80.1% sensitivity). Malekipirbazari and Aksakalli (2015) declared a Random Forests-based classification method as the best classifier compared to FICO (a publicly traded company that produces scoring models most commonly used and distributed by TransUnion, Equifax, and Experian), logistic regression, and support vector machines. Random Forests had the highest accuracy rate, the highest AUC, and the lowest RMSE. Fritzpatrick 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 an artificial neural network is a better alternative than a decision tree and logistic regression when data availability in a dataset is high.

The summary of the methods compared in empirical articles is shown in Table 1. Logistic regression, neural networks, decision trees, discriminant analysis, and support vector machines are the most commonly compared techniques. Neural networks outperform the other methods 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 et al. (2008) | • | • I | | • | | | | | | | • | | | |
| Ince, Aktan (2009) | • | • I | • | • | | | | | | | | | | |
| Tsai et al.(2009) | • | • I | | | | | | | | | | | • I | |
| Yeh, Lien (2009) | • | • I | • | • | | | • | • | | | | | | |
| Kočenda, Vojtek (2011) | • I | | • | | | | | | | | | | | |
| Wang et al. (2011) | • | • | • I | | • | | | | | | | | | |
| Akkoç (2012) | • | • | | • | | | | | | | | | | • I |
| Bekhet, Eletter (2014) | • I | • | | | | | | | | | | | | |
| Abdou et al.(2016) | • | • I | • | | | | | | | | | | | |
| Fritzpatrick, Mues (2016) | • | | • I | | | • | | | | | | • | | |
| Imtiaz, Brimicombe (2017) | • | • $\text{I}^{i}$ | •$\text{I}^{bi}$ | | | | | | | | | | | |
| Serrano-Cinca et al.(2015) | • | | | | | | | | • | | | | | |
| Emekter et al. (2015) | • | | | | | | | | | • | | | | |
| Tsai et al. (2014) | • I | | | | • | • | • | | | | | | | |
| Chang et al. (2015) | • | | | | • | | • I | | | | | | | |
| Malekipirbazari, Aksakalli (2015) | • | | | | • | • I | | • | | | | | | |

Note: • – the technique used in a study, I – the technique stated as the best in a study ($\text{I}^{i}$ – with data imputation, $\text{I}^{bi}$ – without data imputation)

Source: Authors' elaboration.

A more advanced classification technique in the literature is ensemble methods. Ensemble learning is a machine learning paradigm in which multiple learners are trained to solve the same problem (Polikar, 2006). Ensemble learners are usually referred to as base learners. Hansen and Salamon (1990) proved that the generalization performance of a neural network can be improved by using an ensemble of similarly configured neural networks. Schapire (1990) showed that a good classification model can be generated by combining weak classifiers through boosting. The generalization ability of an ensemble is usually much stronger than that of a single learner, which makes ensemble methods very attractive (Dietterich, 1997). According to Windeatt and Ardeshir (2004), two necessary conditions should be met for a good ensemble: Accuracy and Diversity. Some researchers claim that models built using ensemble methods have better predictive power than single classifiers (Lessmann et al., 2015). He et al. (2018) created an original ensemble model based on two types of decision trees, a Random Forest and the XGBoost (Extreme Gradient Boosting) model. It was the most effective classification method compared to single classifiers. In addition, the effectiveness of each model was tested on several datasets. The ensemble model was the best in terms of robustness in the context of changes in data structure. Abellán and Castellano (2017) proved that adding a simple classifier to other more complex ones in an ensemble scheme improves the predictive power.

## 3. Methodology

This section describes three statistical methods used in credit scoring. The first method is the logistic regression model. The second method is neural networks. And the last one is ensemble method, and then, we present the tools to evaluate the quality of classification models .

### 3.1. Logistic regression

Logistic regression is a modeling technique that relates the probability of a binary outcome to a set of predictor variables. Logistic regression is the most commonly used technique in the field of credit scoring – particularly in the classification of customers. Among other things, this model can be used to analyze the probability of timely repayment of a loan and to determine whether the customer belongs to one of two groups – reliable or unreliable borrowers (Matuszyk, 2018). Thus, the goal of a logistic regression model in credit scoring is to determine the conditional probability that a given observation belongs to one of two groups (good and bad customers) given the values of the independent variables of that observation. Bad customers refer to those customers who have defaulted on a loan, and good customers refer to those customers who have repaid a 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 study was estimated on categorized variables modified by using WOE (Weight of Evidence) transformation. WOE allows the transformation of a continuous independent variable into a set of bins based on the similarity of the distribution of the dependent variable, i.e. the number of events. Such an approach solves the problems related to outliers, allows modeling nonlinear relationships using linear models, and gives the opportunity to better interpret the relationships found in the data (Siddiqi, 2006). WOE allows measuring the difference between the analyzed groups (e.g. borrowers who pay their debts on time and borrowers who do not pay their payments) within a given category of the selected variables and is calculated as follows:

$$WOE_i = ln\left(\frac{good_i}{bad_i}\right) \tag{1}$$

where: $i$ – category of a variable, *good*$_i$ – the percentage of clients who repay loans on time for category $i$, *bad*$_i$ – the percentage of clients 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:

$$IV = \sum_{i=1}^{n} (good_i - bad_i) \cdot WOE_i \tag{2}$$

If the IV statistic is less than 0.02, then the predictor is not useful for modeling (separating the Good from the Bad). 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/Bad odds ratio.

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

$$Gini = 1 - \sum_{i=1}^{n} ((cbad_i - cbad_{i-1}) \cdot (cgood_i - cgood_{i-1})) \tag{3}$$

where: $n$ – numer of categories, *cgood*$_i$ – the cumulative percentage of clients who repay loans on time for category $i$, *cbad*$_i$ – cumulative percentage of clients who do not pay their liabilities on time for category $i$.

### 3.2. Neural networks

The idea of neural networks comes from the structure of the human brain. In the human brain, a neuron can send (receive) a signal to (from) other neuron(s).

Similarly, neural networks consist of elements, each of which receives many inputs and produces a single output. Neural networks consist of a large number of simple nodes or neuron elements connected from either a single layer or multiple layers.

An important element of each neuron's activity is the function responsible for the value of the output signal, 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:

$$Y = g\left(\sum_{i=0}^{n-1} w_i y_i - \varphi_h\right) \tag{4}$$

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, $\varphi_h$ – threshold activation level.

Neural networks have many advantages. One of the most important: no assumptions about the statistical distribution of variables and error terms need to be satisfied (Matuszczyk, 2018). However, neural networks also have some disadvantages, including the possibility of a very long learning process, the possibility of instability of behavior in the learning process, the possibility of terminating the action of finding the local minimum (without finding the optimum), poor network transparency – difficulties in interpretation, and the difficulty of finding the cause of the errors that occur.

In our analysis we use a popular type of neural network – the multilayer perceptron (MLP). The MLP network is a unidirectional network that usually has the following structure: an input layer, one (or two) hidden layers consisting of sigmoidal neurons, and an output layer consisting of sigmoidal or linear neurons. The backpropagation algorithm is used for the learning process.

### 3.3. Ensemble methods

In recent years, ensemble methods have become increasingly popular in the context of data modeling. The main goal of using these methods is to increase the performance of classification by combining different classifiers (Abellán and Castellano, 2017). It is possible to combine few models in different ways. One approach aggregates predictions from individual models by using different aggregation functions, such as averaging or voting. Another approach often uses model combination methods such as bagging, boosting, or stacking. The bagging method allows the construction of an ensemble model, where the individual classifiers are based on different subsets generated by sampling with replacement (bootstrap). In the final step, the outputs of the models are aggregated using a particular aggregation function, e.g., majority voting. Boosting is based on a sequence of different classifiers built on a dataset with equal weights for each observation. Predictions are computed and the process starts again, but contains observations with different weights inversely proportional to the accuracy of the predictions. *Stacking* is a multi-layer combination of models. The first layer consists of a few individual classifiers. The predictions obtained from the classifiers in the previous layer are input to the next model (in a subsequent layer). Such a model combination may include a few layers (Raschka and Mirjalili, 2017).

### 3.4. Assessing the quality of classification models

Knowledge of the real state of the repayment process can be compared to the prediction generated by a given model. By testing the model against actual observations, it is possible to evaluate the effectiveness of the model and hence its usefulness for new cases (loans). In the following, we briefly describe how the quality and correctness of classification models can be assessed.

### 3.5. Confusion matrix

In most cases it is not enough to know how often the model being evaluated is wrong. It may be more important to know how often it fails to correctly predict a particular outcome. The confusion matrix provides useful insight into the model's ability to predict a particular group. The matrix contains four possible cases (assuming there are only two categories – positive and negative): *True Positives* (TP) – the number of positive cases correctly classified into the 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 negative cases correctly classified into a negative class.

A variety of performance measures can be derived from the confusion matrix, such as*accuracy* (the ratio of correctly classified cases to all cases), *error rate* (the ratio of misclassified cases to all cases), *positive predictive value* (true-positive rate – the ratio of cases correctly classified as positive to all positive cases), *negative predictive value* (false-positive rate – the ratio of cases classified incorrectly as positive to all negative cases), *sensitivity* (precision – the ratio of cases classified correctly as positive to all cases classified positively), *specificity* (the ratio of cases classified correctly as negative to all negative cases).

Type I error rate is the rate of bad customers categorized as good. A high Type I error rate means that the institution is exposed to credit risk. The Type II error rate (also called $\beta$) is the rate of good customers who are categorized as bad. A high Type II error rate means that the institution is exposed to high business risk over a long period of time, which means that the institution has a restrictive lending policy over a long period of time and may lose its market share.

The given confusion matrix is calculated for the given cut-off point, which is a certain threshold used to determine whether an observation belongs to a certain class.

### 3.6. ROC Curve

A practical tool that facilitates the performance of a classification model at all classification thresholds is the ROC (Receiver Operating Characteristic) analysis. The ROC curve plots the rate of true positives versus false positives at various classification thresholds. On the ROC curve, the x-axis is labeled with the unit minus specificity measure, while the y-axis represents the sensitivity. To plot a ROC curve for a scoring classifier, confusion matrices are calculated for different cut-off points. A reasonable scoring classifier should have its ROC curve completely above the diagonal (the random model line, y=x line passing through (0,0) and (1,1)), which means that the true-positive rate should always be above the false-positive rate.

### 3.7. AUC

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

### 3.8. 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.

$$Gini = \frac{AUC - 0.5}{0.5} = \frac{AUC}{0.5} - 1 = 2 \cdot AUC - 1 \tag{5}$$

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

The higher the Gini coefficient, the better the model.

### 3.9. 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. A higher Kolmogorov-Smirnov statistic value is indicative of a better model. It is defined as:

$$KS = \max_{a\in[L,H]}[abs(F_{m,bad}(a) - F_{n,good}(a))] \tag{6}$$

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. Empirical data and analysis

In this section data and results of empirical research are presented.

### 4.1. Data – source and preparation for modelling

The use of data provided by financial institutions is restricted due to different legal regulations and limitations. Therefore, we used social lending data provided by one of the largest peer-to-peer lending companies in the US. The dataset was downloaded from the website Lending Club [3] for the period between June 2007 and June 2018. We examined 1911592 loans, representing a total loan amount of approximately $ 28.5 billion. We eliminated loans that had not yet been issued or reached maturity to include cases with “paid in full” or “defaulted” status. The final sample includes 119,160 observations from 2011 to 2013. Loans funded through 2013 are analyzed because the status of later loans (defaulted or non-defaulted) is still unknown because the maturity of these loans is 36 or 60 months. For example, the status of a 36-month loan funded in September 2014 may not be known until September 2019. In this article, the 2-year time frame is chosen (July 1, 2011 – July 30, 2013). In credit scoring, the 12-month time frame is commonly used in model development, but the time window can be extended to 24 months (Matuszyk, 2018). The data examined included credit records with all credit information commonly used to assign a score, so financial and other borrower-specific characteristics were available.

The objective of credit scoring is to determine the conditional probability of default of a given observation given the values of the independent variables. An important step in developing the predictive model is then to define the response variable that divides all observations into two separate groups: good and bad customers. In this research, we aim to analyze customers who repay the loan and those who default on a loan. We based the definition of good and bad customers on a variable that describes the loan status. Thus, fully repaid loans are interpreted as good customers (marked as 0) and loans that have been written off are considered bad customers (marked as 1).

The next step was to select characteristics with good predictive power for the response variable. After the analysis, the details of which are described below, we selected 17 of these features to be used in the predictive modeling.

Originally, the dataset we used contained more than 100 variables. Variables with many missing values were then eliminated (at least 90% of missing values). We then removed technical variables (e.g., links to websites, observation IDs), features that are difficult to use in the modeling process (e.g., unstructured job description provided by the borrower), and variables that are not known at the time of the loan application – there is no such information in the document submitted by the borrower (e.g., rate amount, credit risk class, interest rate, last payment amount).

Some of the remaining variables were difficult to use in the modeling in their original form, e.g. comment about the loan delivered by the borrower or date of opening of the first loan product by the borrower. Due to this fact, new variables were created (based on the original variables): number of letters in the credit comment (there is a suspicion that individuals who add a very long comment as a motivation for needing credit are more risky because they want to get the credit at any cost), number of years from credit release (assuming that the date of credit release is almost identical to the date of submission of the credit application by a borrower).

In the next step of model building, the dataset was constrained to include only the most predictive variables. To select the variables, the fine classification procedure was used. This is a binning procedure. In this process, the independent variable is divided into quantile groups. And then for each group, there are several good customers and several bad customers. Based on this, the following measures are calculated: the bad rate (shows the proportion of observations that have a value of 1 for the target variable compared to all records in the given group) and the WOE (Weight of Evidence). It is useful to calculate the measures of predictive power of the variable, such as the Gini coefficient and IV (Information Value). These indicators are very helpful in variable selection. However, before the final variable selection, the correlation analysis was performed. Kendall’s tau correlation coefficient was used as a measure of the relationship between ordinal variables (obtained with WOE transformation). After correlation analysis, the final variable selection was based on the Gini coefficient. Variables with Gini coefficient higher than 5% were considered as the most predictive variables. According to this rule, the final set of variables includes 16 explanatory variables. The characteristics used in our predictive models are described in Table 2.

Table 2: List of final features used in the estimation process

| Variable's name | Variable's description | Gini coefficient |
|---|---|---|
| *Term* | The number of payments (in months): 36 months (78%), 60 months (22%). | 0.17113 |
| *Acc_open_past_24mths* | The number of trades opened in past 24 months | 0.11715 |
| *Dti* | A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested loan, divided by the borrower's self-reported monthly income | 0.11669 |
| *Revol_util* | Revolving line utilization rate | 0.10943 |
| *Annual_inc* | The self-reported annual income provided by the borrower during registration | 0.10906 |
| *Percent_bc_gt_75* | Percentage of all bankcard accounts greater than 75% of limit | 0.10460 |
| *Verification_status* | Indicates if income was verified: verified (45.5%), not verified (34.4%), source verified (20.1%) | 0.10199 |
| *Inq_last_6mths* | The number of inquiries in past 6 months: 0 (50%), 1 (28.3%), 2 (13.7%), 3+ (8%) | 0.10078 |
| *Loan_amnt* | The amount of the loan | 0.09831 |
| *Num_tl_op_past_12m* | The number of accounts opened in past 12 months | 0.09365 |
| *Purpose* | A category provided by the borrower for the loan request: 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%) | 0.09249 |
| *Mo_sin_rcnt_tl* | Months since most recent account opened | 0.08401 |
| *Mths_since_recent_bc* | Months since most recent bankcard account opened | 0.07545 |
| *Num_rev_tl_bal_gt_0* | The number of revolving trades with balance greater than 0 | 0.07218 |
| *Mort_acc* | The number of mortgage accounts | 0.05926 |
| *Earliest_cr_line_n* | The number of months from the date that the borrower's earliest reported credit line was opened (modified original variable) | 0.05216 |

Source: Authors' calculations.

Descriptive statistics of continuous variables are shown in Table 3.

Table 3: Descriptive statistics of continuous variables

| Variable/statistic | Average | Standard deviation | Minimum | Median | Maximum | Kurtosis | Missing percentage |
|---|---|---|---|---|---|---|---|
| *Acc_open_past_24mths* | 3.91 | 2.68 | 0.00 | 3.00 | 40.00 | 3.19 | 16.80 |
| *Dti* | 16.67 | 7.59 | 0.00 | 16.44 | 34.99 | 0.63 | 0.00 |
| *Revol_util* | 58.08 | 24.00 | 0.00 | 60.90 | 122.50 | -0.57 | 0.10 |
| *Annual_inc* | 71,546 | 55,790 | 4,800 | 61,000 | 71,41,778 | 3,875 | 0.00 |
| *Percent_bc_gt_75* | 34.64 | 27.53 | 0.00 | 33.00 | 94.00 | -1.32 | 17.50 |
| *Loan_amnt* | 14,046 | 8,168 | 1,000 | 12,000 | 35,000 | -0.10 | 0.00 |
| *Num_tl_op_past_12m* | 1.82 | 1.57 | 0.00 | 2.00 | 25.00 | 4.77 | 33.80 |
| *Mo_sin_rcnt_tl* | 8.95 | 9.67 | 0.00 | 6.00 | 99.00 | 13.80 | 33.80 |
| *Mths_since_recent_bc* | 22.14 | 21.69 | 0.00 | 14.00 | 99.00 | 1.17 | 17.40 |
| *Num_rev_tl_bal_gt_0* | 5.75 | 2.96 | 0.00 | 5.00 | 37.00 | 2.36 | 33.80 |
| *Mort_acc* | 1.74 | 2.19 | 0.00 | 1.00 | 29.00 | 3.41 | 16.80 |
| *Earliest_cr_line_n* | 14.59 | 6.93 | 3.00 | 13.00 | 62.00 | 1.78 | 0.00 |

Source: Authors' calculations.

Additionally, Table 4 presents the results of independence Chi$^2$ 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 |
|---|---|---|
| *Term* | 2719.3 | 0.0000 |
| *Verification status* | 597.12 | 0.0000 |
| *Inq_last_6mths* | 585.46 | 0.0000 |
| *Purpose* | 598.11 | 0.0000 |

Source: Authors' calculations.

According to the results from Table 4, the null hypothesis about the independence of variables is rejected for all variables (p-value is smaller than the assumed 5% significance level in each case), so there are statistically significant dependencies between the target variable and certain explanatory variables.

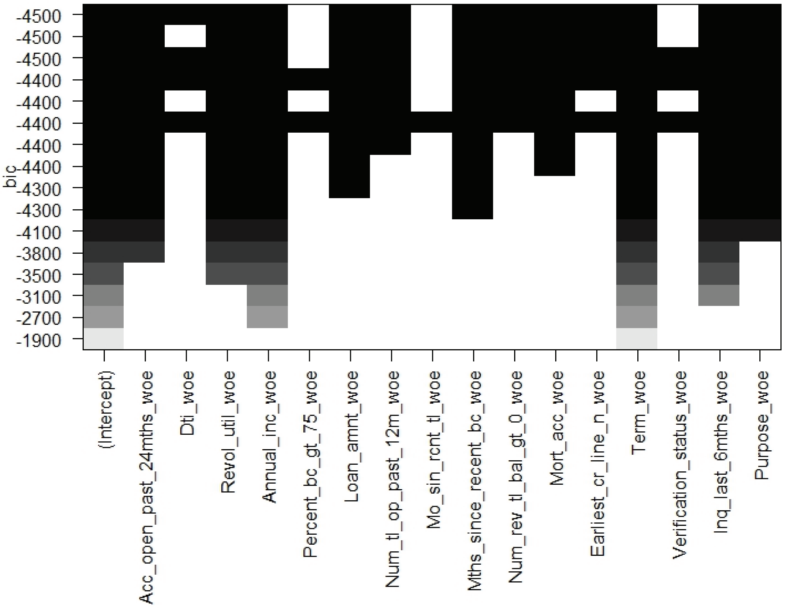
The sample analyzed consists of 119,160 credit records. Of these, 70% were randomly selected as a training sample to estimate parameters of the credit scoring model. The remaining 30% were retained for validation. Both subsets contain a similar proportion of delinquencies, about 16%.

### 4.2. Logistic regression model

The weight-of-evidence method (WOE) with a coarse classification procedure was applied to transformation variables. Coarse classification was used to create fewer categories by merging similar adjacent groups. The graphical results of the coarse classification procedure are presented in the appendix of this paper.

The selection of independent variables in a logistic regression model posed a challenging problem. This task was approached using techniques such as the backward stepwise regression method and the forward stepwise regression method. In addition, graphical visualization was used to display the Bayesian Information Criterion (BIC) values for models with different sets of explanatory variables (Figure 1).

Figure 1: Values of BIC statistics for models with different sets of explanatory variables



Note: 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.

Source: Authors' calculations.

Results of 6 best models are presented in Table 5.

Table 5: Results of logistic models

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

Note: ***,**, * donate the significance level of 1%, 5%, 10% respectively. Standard errors are reported in the parenthesis.

Source: Authors' calculation.

The final model was selected based on the Akaike information criteria (AIC) and the Schwartz information criteria (BIC). The information criteria consider the goodness of fit of the model and the simplicity of the model. The model with the lowest information criteria is preferred. The AIC indicates that the best model is the model number 2, while the BIC states that the best model is the model number 4. However, the literature shows that the AIC criterion tends to consider a model with too many parameters as the best. This is consistent with the analyzed case – model 2 contains more parameters compared to model 4. Therefore, model 4 was finally considered as the best logistic regression model.

The diagnostic tests for logistic regression model 4 were then conducted. The goodness of fit of the model to the data is tested using the Hosmer-Lemeshow test. The null hypothesis in this test is a statement of goodness of fit of the model to the data. In all the versions of this test performed (Hosmer-Lemeshow (5 bins), chi2 = 3.0610, *p-value* = 0.3823; Hosmer-Lemeshow (10 bins), chi2 = 10.8666, *p-value* = 0.2094; Hosmer-Lemeshow (15 bins), chi2 = 20.10210, *p-valu*e = 0.0927), there is no reason to reject the null hypothesis at the 5% significance level, therefore the model is well fitted to the data. However, this test has many drawbacks, including high sensitivity to the number of bins. For this reason, there are some critical opinions about this test (Allison, 2013). For this reason, another test is performed – the Osius-Rojek test, which also checks the goodness of fit of the model. According to the test (*p-valu*e = 0.7426), there is no reason to reject the null hypothesis about goodness of fit of the model to the data at 5% level of significance. Thus, the model specification can be considered as the correct one. In addition, the likelihood ratio test (LR) is performed to test the joint significance of all variables in the model. According to the results of this test (*p-value* = 0.0000), the null hypothesis about the lack of significance of all variables at the 5% significance level is rejected, therefore, it is assumed that the variables in the selected model are significant.

The results of the tests described above were as expected, therefore model 4 is considered as the best choice. The next step was to perform the prediction. For all observations in the training set as well as in the validation set, the failure probability estimated by the model was assigned. The evaluation of the model prediction is shown in Table 6.

Table 6: The 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: Authors' calculations.

### 4.3. Neural networks model

The scale and distribution of variables may differ. Differences in the scales of the input variables can increase the difficulty of the problem being modeled. For example, large input values (e.g., measured in thousands of units) can lead to a model that learns large weight values (Szeliga, 2017). In turn, a model with large weight values is often unstable. Also, a target variable with a large spread of values makes the learning process unstable, as it leads to weight values changing dramatically. In practice, it is almost always beneficial to apply preprocessing transformations to variables before training a neural network model. Scaling data is useful to improve neural network stability and modeling performance. In this process, the values of variables are rescaled so that the minimum value is 0 and the maximum value is 1. By transforming the inputs in this way, training can be faster and the probability of getting stuck in local optima can be reduced. We used the original values of a variable (not after the WOE transformation) transformed by subtracting the minimum and dividing by the range (the difference between the largest and smallest values).

We use a multilayer perceptron neural network. An important point in constructing this type of neural network is to determine the number of layers and the number of neurons in each layer. In this paper, a neural network with only one hidden layer is used. It is claimed that a multilayer perceptron neural 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 more easily or conveniently – the more complex neural network (in terms of its structure) can give better results. In practice, it is sufficient 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 is rarely used in practice (Hastie et al., 2008).

The next issue that arises is the selection of the number of neurons in the hidden layer. The selection of the number of neurons in the hidden layer is a very important part of the overall neural network architecture. Although hidden layers do not directly interact with the external environment, they affect the final output. There are different approaches to find out a large number of hidden nodes in the hidden layer. The rule of thumb is that the number of hidden neurons should be in the range between the size of the input layer and the size of the output layer (Panchal and Panchal, 2014). The try-and-error method assumes repeated trials. In the forward approach, a small number of hidden neurons, usually two, are started, then trained and tested. In the next step, the number of hidden neurons is gradually increased. The process is repeated until the test results do not improve.

In our research, we combined these two approaches. We started with two hidden neurons in the hidden layer. The neural network was trained and evaluated using the AUC measure. Based on the rule of thumb mentioned above, we repeated the process until the variant with 34 hidden neurons in the hidden layer. The AUC measure was then used as a criterion for selecting the best network structure. Table 7 shows the AUC values for network structures with different numbers of neurons in the hidden layer.

Table 7: AUC values depending on a certain number of neural neurons in the hidden layer

| Number of neurons | AUC |
|---|---|
| 2 | 0.66533 |
| 3 | 0.66632 |
| 4 | 0.66621 |
| 5 | 0.66891 |
| 6 | 0.66856 |
| 7 | 0.66905 |
| 8 | 0.67048 |
| 9 | 0.67051 |
| 10 | 0.66943 |
| 11 | 0.67136 |
| 12 | 0.67061 |
| 13 | 0.67141 |
| 14 | 0.67092 |
| 15 | 0.67193 |
| 16 | 0.67289 |
| 17 | 0.6732 |
| 18 | 0.66898 |
| 19 | 0.67261 |
| 20 | 0.67251 |
| 21 | 0.67239 |
| 22 | 0.67259 |
| 23 | 0.67219 |
| 24 | 0.67137 |
| 25 | 0.67159 |
| 26 | 0.67201 |
| 27 | 0.67132 |
| 28 | 0.67297 |
| 29 | 0.67315 |
| 30 | 0.67254 |
| 31 | 0.67335 |
| 32 | 0.67364 |
| 33 | 0.67330 |
| 34 | 0.67361 |

Source: Authors' calculations.

Results presented in Table 7 show that the best neural network (in terms of AUC measure calculated on the training set) has 32 neurons in the hidden layer. Then a prediction was done and measures of model classification were calculated both for the training and validation set (Table 8).

Table 8: 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: Authors' calculations.

### 4.4. Models built by using ensemble methods

Based on the assumption that combining different techniques in one predictive model could provide better predictive results, two other models were built using ensemble methods. They combined in different ways the two best models from the previous sections.

The first ensemble model combined logistic regression and neural networks in a parallel way by averaging the probabilities obtained from the individual models. The evaluation of the model prediction is shown in Table 9.

Table 9: Evaluation of model prediction – the first ensemble model

| Measure | Training set | Validation set |
|---|---|---|
| AUC | 0.6823 | 0.6732 |
| Gini | 0.3646 | 0.3464 |
| KS | 0.2629 | 0.2573 |

Source: Authors' calculations.

The second ensemble model combined the individual models in the other way – instead of using an average function, logistic regression was used to produce the final output. Probabilities returned from the logistic regression and the neural network were inputs to the logistic regression, which produced the final probability values. The evaluation of the model prediction is shown in Table 10.

Table 10: Evaluation of model prediction – the second ensemble model

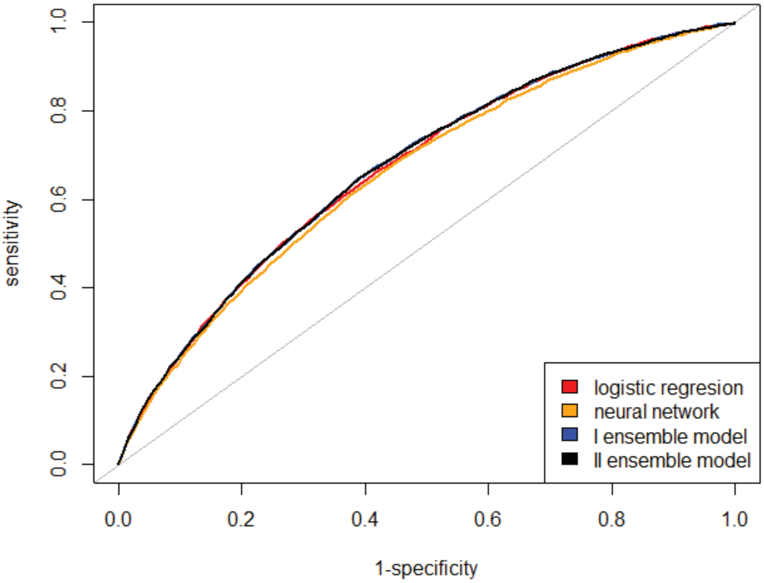
| Measure | Training set | Validation set |
|---|---|---|
| AUC | 0.6821 | 0.6727 |
| Gini | 0.3642 | 0.3453 |
| KS | 0.2632 | 0.2570 |

Source: Authors' calculations.

## 5. Results and discussion

The next step of the analysis was to compare the performance of all models. It was checked which model was better in terms of prediction. We started by graphically comparing the classification abilities of the models based on the ROC curves for the validation set (Figure 2).

Figure 2: ROC curves for final models (validation set)



Source: Authors' calculations.

Based on the ROC curves graph, it is very hard to decide which model is the best one. It can be noticed that probably neural network model in the worst way classifies borrowers into two groups, good customers and bad customers, as the neural network's ROC curve line is the least shifted towards the point (1.1) in comparison to the other models' ROC curves.

To evaluate the overall credit scoring capability of the proposed credit scoring models, performance across all possible classification thresholds (AUC), Gini coefficient, and the maximum difference between true positive rate and false positive rate obtained for different cut-off points (Kolmogorov-Smirnov statistic) are used. Moreover, accuracy, error ratio, positive predictive value (PPV), and negative predictive value (NPV) are computed. The following Table 11 shows the evaluation measures of the two credit scoring models calculated on a validation set.

Table 11: Evaluation of models' classification (a validation set; in %)

| Measure/ model | Logistic regression | | Neural network | | I ensemble model | | II ensemble model | |
|---|---|---|---|---|---|---|---|---|
| | *Z**=25 | *Z**=50 | *Z**=25 | *Z**=50 | *Z**=25 | *Z**=50 | *Z**=25 | *Z**=50 |
| AUC | 67.08 | | 66.01 | | 67.32 | | 67.27 | |
| Gini | 34.16 | | 32.02 | | 34.64 | | 34.53 | |
| KS | 24.47 | | 23.59 | | 25.73 | | 25.70 | |
| Accuracy | 78.71 | 84.07 | 76.14 | 84.11 | 77.87 | 84.07 | 78.55 | 83.9 |
| Error ratio | 21.29 | 15.93 | 23.86 | 15.89 | 22.13 | 15.93 | 21.45 | 16.1 |
| PPV | 31.07 | 48.91 | 28.62 | 53.05 | 30.35 | 47.06 | 31.09 | 44.05 |
| NPV | 86.58 | 84.16 | 86.97 | 84.25 | 86.78 | 84.14 | 86.67 | 84.51 |

Notes: Some measures depending on cut-off points (marked as Z*).

Source: Authors' calculations.

As the results in Table 11 show, the I ensemble model (a combination of logistic regression and neural network by averaging the probabilities obtained from both models) has higher AUC, higher Gini coefficient, and higher Kolmogorov-Smirnov statistic compared to other models. Therefore, we can conclude that the I ensemble model allows to successfully reduce the potential risks of losses due to misclassification costs.

Since the I-ensemble model and the II -ensemble model have similar discriminatory power, the formal test (on the validation set) is performed to check whether the ROC -curves for the compared models are equally good (the null hypothesis assumes that the ROC -curves are equally good). Based on the obtained test results, the null hypothesis is rejected at 5% significance level (*p-value* = 0.000). Thus, there are statistically significant differences in the classification effectiveness of the analyzed models.

Analyzing the readings from Table 11, we can see that the choice of a cut-off point (which defines a boundary between bad and good cases) is extremely important. When the cut-off point is increased, the accuracy and precision of the positive prediction also increase and consequently the classification error and precision of the negative prediction decrease. This is a key element in a business environment, especially in financial institutions, as it relates to the control of the sales process (loans granted) as well as to the quality of the loan portfolio (a certain percentage of customers who do not pay their liabilities according to the contract). For this reason, an econometric model that has a very good predictive power is a very important and practical business tool.

To summarize. When comparing logistic regression and neural networks, we found that logistic regression performs better than neural network in classifying customers as good or bad (similar conclusions were found by Bekhet and Eletter, 2014). The reason for our result may be the use of transformation WOE in constructing a logistic regression model, which certainly improved its results and has not always been used by other researchers in the context of comparing classifiers. In addition, the poor efficiency of neural networks could result from the implementation of a rather simple network structure which, although widely used, could prove less effective than more complex network structures, including more hidden layers.

The proposed ensemble model (a combination of logistic regression and neural network by averaging the probabilities obtained from both models) showed higher AUC, Gini coefficient and Kolmogorov-Smirnov statistics compared to other models. Therefore, we can conclude that the ensemble model allows to successfully reduce the potential risks of losses due to misclassification costs. Further potential research in this area could address the application of more advanced features in ensemble methods.

## 6. Conclusions

Financial institutions engaged in the activity of money lending play an important role in everyone's life because they improve the quality of life. Moreover, the state of financial institutions affects the functioning of the finances of the whole country. Recently, an alternative way of raising money – social loans – is gaining quite significant importance. An important element of all financial institutions engaged in lending activity is strict control of credit risk (its minimization), which affects the financial condition of the institution.

In the era of growing importance of data and advanced data analysis, a number of methods are available to the companies to optimize the management of various processes. Different types of econometric models are used to measure and manage credit risk properly. Many studies compare the effectiveness of different techniques to properly classify good customers (on-time loan repayments) and bad customers (not on-time loan repayments). The choice of the best classifier plays a key role in controlling the proportion of non-performing loans in the portfolio of a given financial institution, which affects its functioning.

Our paper reviews the literature related to a comparison of classifier results. The two most commonly compared techniques are the logistic regression model and the neural network model. In our study, we analyze these two techniques on real social credit data. The aim of the study was to test the hypothesis whether the neural network model is a better classifier compared to logistic regression and whether ensemble methods have better performance than base learners (regression analysis and neural network).

Our results show that logistic regression performs better than neural networks. When comparing logistic regression and neural networks, we found that logistic regression performs better than neural networks in classifying customers as good or bad. Combining logistic regression and neural network by averaging the probabilities obtained from both models suggests that the ensemble model successfully reduces the potential risks of losses due to misclassification costs. Further potential research in this area could be devoted to the application of more advanced features in ensemble methods.

## References

Abdou, H.A., Pointon, J., El-Masry, A. (2008) "Neural nets versus conventional techniques in credit scoring in Egyptian banking", *Expert Systems with Applications*, Vol. 35, No.3, pp. 1275–1292, doi: 10.1016/j.eswa.2007.08.030.

Abdou, H.A. et al. (2016) "Predicting creditworthiness in retail banking with limited scoring data", *Knowledge-Based Systems*, Vol. 103, pp. 89–103, doi: 10.1016/j.knosys.2016. 03.023.

Abellán, J., Castellano, J.G. (2017) "A comparative study on base classifiers in ensemble methods for credit scoring", *Expert Systems with Applications*, Vol. 73, pp. 1–10, doi: 10.1016/j.eswa.2016.12.020.

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*, Vol. 222, No. 1, pp. 168–178, doi: 10.1016/j.ejor.2012.04.009.

Allison, P. (2013) "Why I Don't Trust the Hosmer-Lemeshow Test for Logistic Regression", *Statistical Horizons*, https://statisticalhorizons.com/hosmer-lemeshow, [Accessed: February 26, 2021].

Bekhet, H., Eletter, S. (2014) "Credit risk assessment model for Jordanian commercial banks: Neural scoring approach", *Review of Development Finance*, Vol. 4, No. 1, pp. 20–28, doi: 10.1016/j.rdf.2014.03.002.

Chang, S., Kim, S.D., Kondo, G. (2015) „Predicting Default Risk of Lending Club Loans", *Machine Learning*, CS229, pp. 1–5.

Dietterich, T.G. (1997) "Machine-learning research", *AI magazine*, Vol. 18, No. 4, pp. 97–136, doi: 10.1609/aimag.v18i4.1324.

Emekter, R. et al. (2015) "Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending", *Applied Economics*, Vol. 47, No. 1, pp. 54–70, doi: 10.1080/00036846. 2014.962222.

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*, Vol. 249, No. 2, pp. 427–439, doi: 10.1016/j.ejor. 2015.09.014.

Hand, D.J., Henley, W.E. (1997) "Statistical classification methods in consumer credit scoring: a review", *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, Vol. 160, No. 3, pp. 523–541, doi: 10.1111/j.1467-985X.1997.00078.x.

Hansen, L.K., Salamon, P. (1990) "Neural network ensembles", *IEEE transactions on pattern analysis and machine intelligence*, Vol. 12, No. 10, pp. 993–1001, doi: 10.1109/34.58871.

Hastie, T., Tibshirani, R., Friedman, J. (2008) "The Elements of Statistical Learning – Data Mining, Inference, and Prediction", Springer, second edition.

He, H., Zhang, W., Zhang, S. (2018) "A novel ensemble method for credit scoring: Adaption of different imbalance ratios", *Expert Systems With Applications*, Vol. 98, pp. 105–117, doi: 10.1016/j.eswa.2018.01.012.

Imtiaz, S., Brimicombe, A. (2017) "A Better Comparison Summary of Credit Scoring Classification", *International Journal of Advanced Computer Science and Applications*, Vol. 8, No. 7, pp. 1–4, doi: 10.14569/IJACSA.2017.080701.

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*, Vol. 10, No. 3, pp. 233–240, doi: 10.3846/1611-1699.2009.10.233-240.

Kočenda, E., Vojtek, M. (2011) "Default Predictors in Retail Credit Scoring: Evidence from Czech Banking Data" *Emerging Markets Finance and Trade*, Vol. 47, No. 6, pp. 80–98, doi: 10.2753/REE1540-496X470605.

Lahsasna, A., Ainon, R.N., Teh, Y.W. (2010) "Credit Scoring Models Using Soft Computing Methods: A Survey", *Int. Arab J. Inf. Technol.*, Vol. 7, No. 2, pp. 115–123.

Lee, T.S. et al. (2002) "Credit scoring using the hybrid neural discriminant technique", *Expert Systems with applications*, Vol. 23, No. 3, pp. 245–254, doi: 10.1016/S0957-4174(02)00044-1.

Lessmann, S. et al. (2015) "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research", *European Journal of Operational Research*, Vol. 16, No. 1, pp. 124–136, doi: 10.1016/j.ejor.2015.05.030.

Malekipirbazari, M., Aksakalli, V. (2015) "Risk assessment in social lending via random forests", *Expert Systems with Applications*, Vol. 42, No. 10, pp. 4621–4631, doi: 10.1016/j.eswa.2015.02.001.

Matuszyk, A. (2018) "Credit Scoring", CeDeWu, second edition.

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*, Vol. 3, No. 11, pp. 455–464.

Polikar, R. (2006) "Ensemble based systems in decision making", *IEEE Circuits and systems magazine*, Vol. 6, No. 3, pp. 21–45, doi: 10.1109/MCAS.2006.1688199.

Raschka, S., Mirjalili,V. (2017) “Python Machine Learning”, Packt Publishing Ltd.

Schapire, R.E. (1990) “The strength of weak learnability”, *Machine learning*, Vol. 5, No. 2, pp. 197–227, doi: 10.1007/BF00116037.

Serrano-Cinca, C., Gutiérrez-Nieto, B., López-Palacios, L. (2015) “Determinants of Default in P2P Lending”, *PLoS ONE*, Vol. 10, No. 10, pp. 1–22, doi: 10.1371/journal.pone.0139427.

Siddiqi, N. (2006) “Credit risk scorecards: developing and implementing intelligent credit scoring”, Vol. 3, John Wiley & Sons.

Szeliga, M. (2017) „Data science i uczenie maszynowe”, PWN, first edition.

Tsai, K., Ramiah, S., Singh, S. (2014) “Peer Lending Risk Predictor”, CS229, pp. 1–5.

Tsai, M.C. et al. (2009) “The consumer loan default predicting model – An application of DEA-DA and neural network”, *Expert Systems with Applications*, Vol. 36, No. 9, pp. 11682–11690, doi: 10.1016/j.eswa.2009.03.009.

Wang, G. et al. (2011) „A comparative assessment of ensemble learning for credit scoring”, *Expert Systems with Applications*, Vol. 38, No. 1, pp. 223–230, doi: 10.1016/j.eswa. 2010.06.048.

West, D. (2000) “Neural network credit scoring models”, *Computers & Operations Research*, Vol. 27, No. 11-12, pp. 1131–1152, doi: 10.1016/S0305-0548(99)00149-5.

Windeatt, T., Ardeshir, G. (2004) “Decision tree simplification for classifier ensembles”, *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 18, No. 5, pp. 749–776, doi: 10.1142/S021800140400340X.

Wu, D., Olson, D.L. (2010) “Enterprise risk management: coping with model risk in a large bank”, *Journal of the Operational Research Society*, Vol. 61, No. 2, pp. 179–190, doi: 10.1057/jors.2008.144.

Yeh, I.C., 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*, Vol. 36, No. 2, pp. 2473–2480, doi: 10.1016/j.eswa.2007.12.020.

# Primjena ansambl metoda, logističke regresije i neuronske mreže na mogućnost predviđanja Peer-to-Peer pozajmljivanja

*Aneta Dzik-Walczak*[1], *Mateusz Heba*[2]

### *Sažetak*

*Procjena kreditne sposobnosti postaje izuzetno važna s obzirom na sve intenzivniju konkurenciju među financijskim institucijama tako da čak i neznatno unapređivanje točnosti predviđanja može rezultirati značajnom uštedom. Financijske institucije traže optimalne strategije pomoću modela procjene kreditne sposobnosti. Stoga je proučavanje alata za procjenu kreditne sposobnosti široko rasprostranjeno. Kao rezultat toga, razvijene su različite parametarske statističke metode, ne-parametarski statistički alati i pristupi programskom računanju kako bi se povećala točnost modela procjene kreditne sposobnosti. U ovom radu primjenjuju se različiti pristupi za klasifikaciju kupaca, kao onih koji vraćaju zajam i onih koji ne mogu podmirivati svoje obveze. Svrha ove studije je istražiti uspješnost dviju tehnika vrednovanja kreditne sposobnosti, modela logističke regresije, procijenjene na temelju kategorizirane varijable modificirane pomoću WOE (Weight of Evidence) transformacije, i neuronskih mreža. Nadalje, istražuje se da li kombiniranje više klasifikatora i testiranje prikupljenih informacija ansambl metodom doprinosi boljim rezultatima. Da bi se procijenila izvedivost i učinkovitost ovih metoda, provodi se analiza podataka Lending Cluba. Istražuje se P2P pozajmljivanje, odnosno uzajamno pozajmljivanje bez posredovanja financijskih institucija, koje se još naziva i socijalno pozajmljivanje. Na temelju provedenog istraživanja, može se zaključiti da model logističke regresije daje bolje rezultate od neuronskih mreža. Izgleda da je predloženi ansambl model (kombinirajući logističku regresiju i neuronsku mrežu s prosjekom vjerojatnosti dobivenih iz oba modela) imao veću AUC krivulju, Gini koeficijent i Kolmogorov-Smirnov test veću statističku vrijednost u usporedbi s drugim modelima. Stoga možemo zaključiti da ansambl model omogućuje uspješno reduciranje mogućih rizika od gubitaka koji nastaju uslijed pogrešne klasifikacije troškova.*

**Ključne riječi:** *procjena kreditne sposobnosti, ansambl metode, logistička regresija, neuronske mreže, P2P pozajmljivanje/ uzajamno pozajmljivanje*

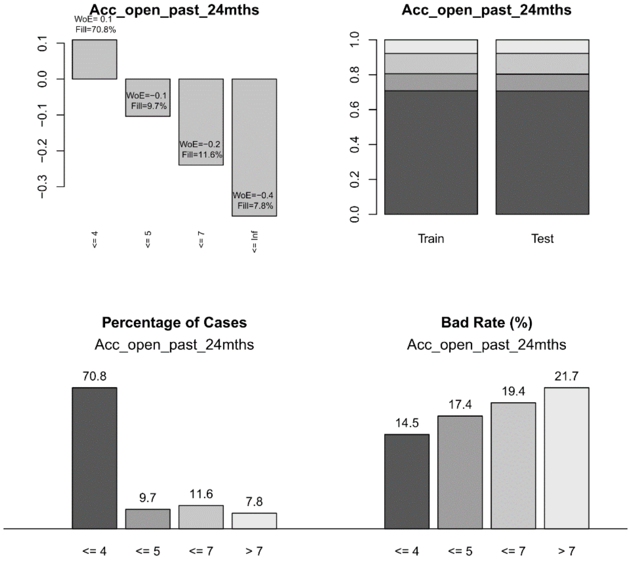
***JEL klasifikacija:*** *G21, G32*

[1] *Docentica, University of Warsaw – Faculty of Economic Sciences, Długa 44/50, 00-241 Varšava, Poljska. Znanstveni interes: ekonometrijske metode i modeli, korporativne financije. Tel.: +48 22 55 49 111. Fax: 22 831 28 46. E-mail: adzik@wne.uw.edu.pl. Osobna web stranica: http://www.wne.uw.edu.pl/index.php/pl/profile/view/155/.*

[2] *Doktorand, University of Warsaw – Faculty of Economic Sciences, Długa 44/50, 00-241 Varšava, Poljska. Znanstveni interes: ekonometrijske metode i modeli, financijski rizik, kreditni rizik, korporativne financije. Tel.: +48 22 55 49 111. Fax: 22 831 28 46. E-mail: mheba@wne.uw.edu.pl. Osobna web stranica: https://www.wne.uw.edu.pl/index.php/pl/profile/ view/336/.*

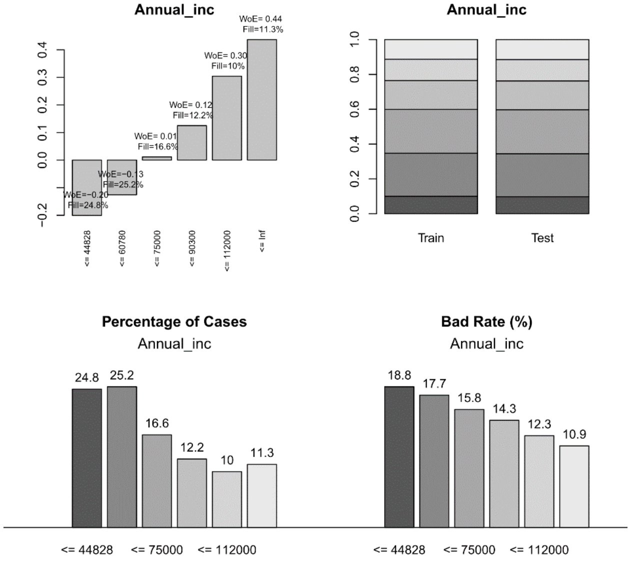
## Appendix – *Coarse classing* procedure results for particular variables

Figure 3: Variable *Acc_open_past_24mths – coarse classing*



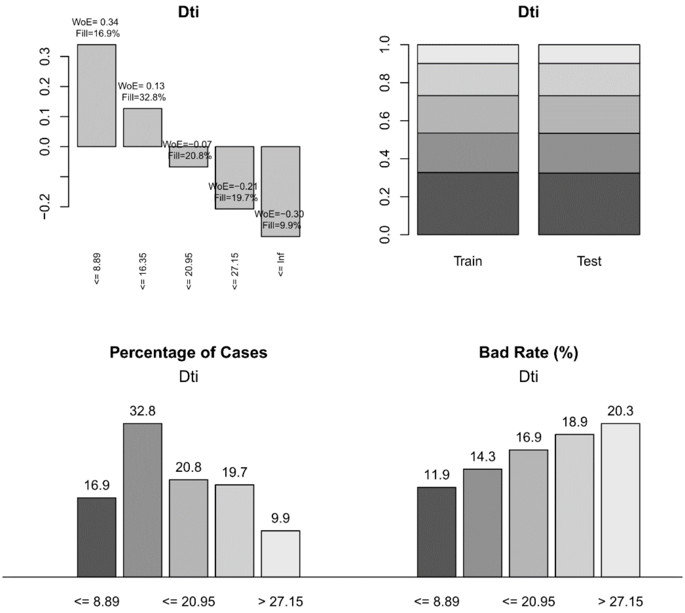
Source: Authors' calculations.

Figure 4: Variable *Annual_inc – coarse classing*



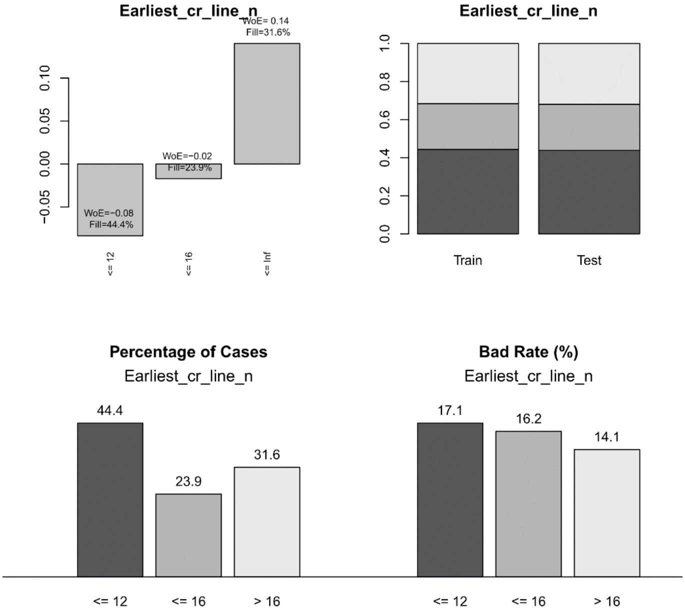
Source: Authors' calculations.

Figure 5: Variable *Dti – coarse classing*



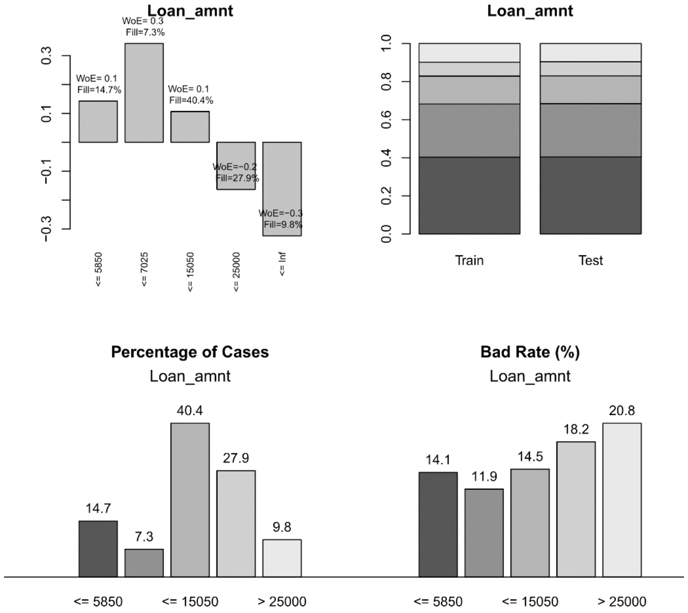
Source: Authors' calculations.

Figure 6: Variable *Earliest_cr_line_n – coarse classing*



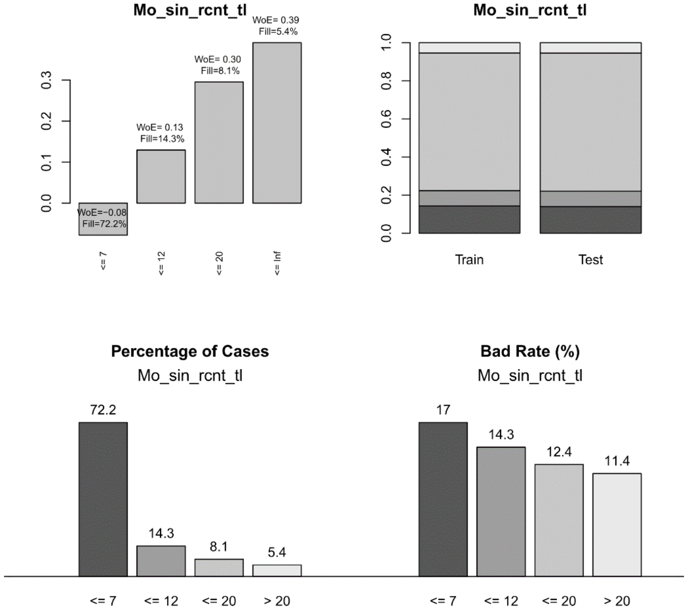
Source: Authors' calculations.

Figure 7: Variable *Loan_amnt – coarse classing*



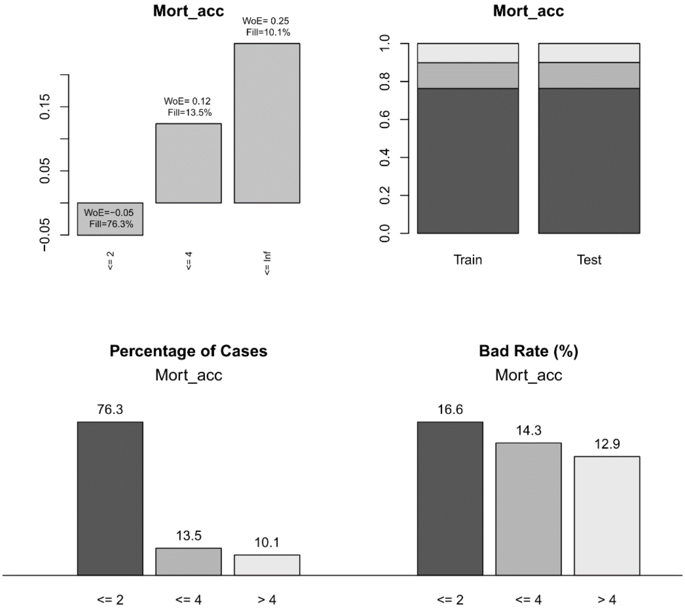
Source: Authors' calculations.

Figure 8: Variable *Mo_sin_rcnt_tl – coarse classing*



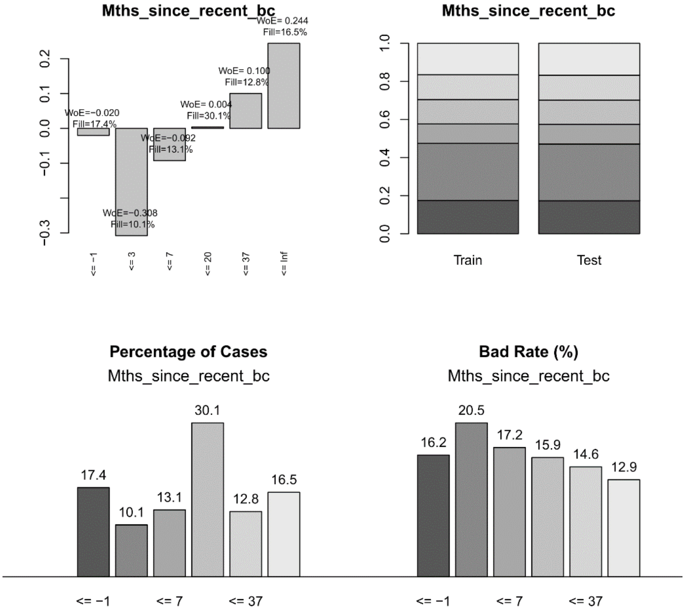
Source: Authors' calculations.

Figure 9: Variable *Mort_acc – coarse classing*



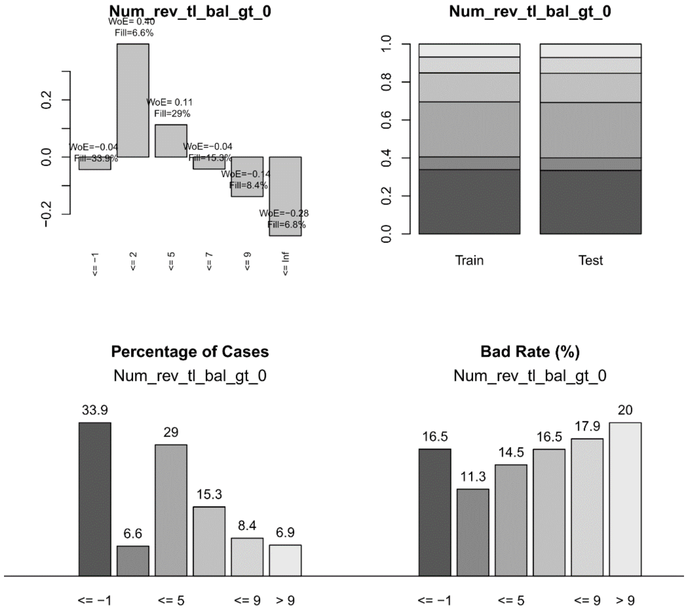
Source: Authors' calculations.

Figure 10: Variable *Mths_since_recent_bc – coarse classing*



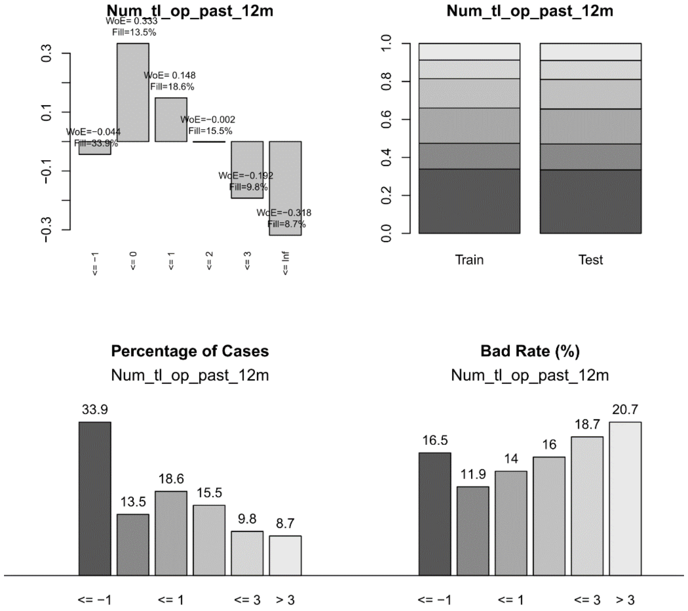
Source: Authors' calculations.

Figure 11: Variable *Num_rev_tl_bal_gt_0 – coarse classing*



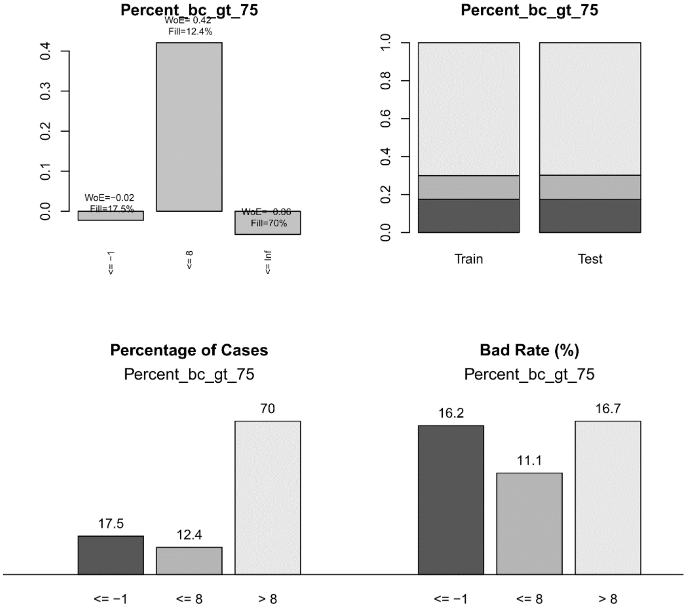
Source: Authors' calculations.

Figure 12: Variable *Num_tl_op_past_12m – coarse classing*



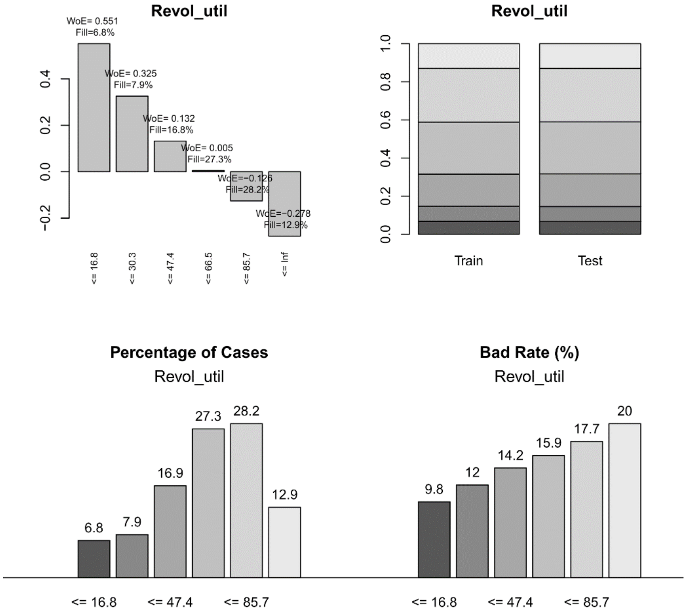
Source: Authors' calculations.

Figure 13: Variable *Percent_bc_gt_75 – coarse classing*



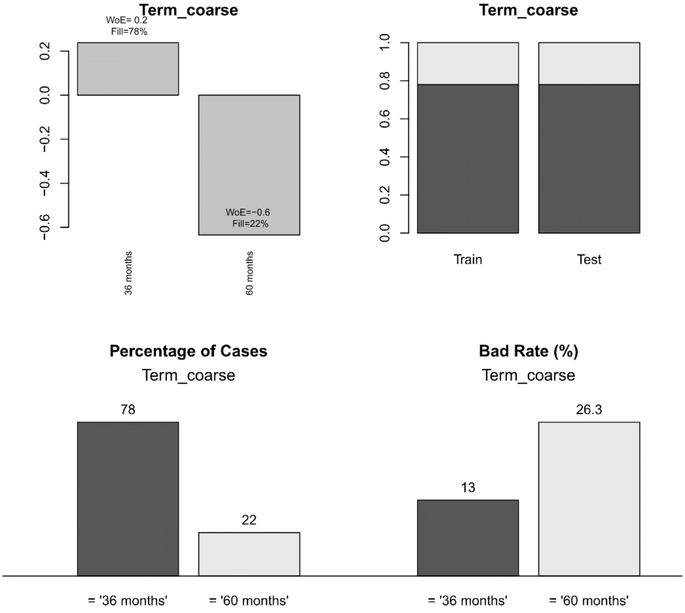
Source: Authors' calculations.

Figure 14: Variable *Revol_util – coarse classing*



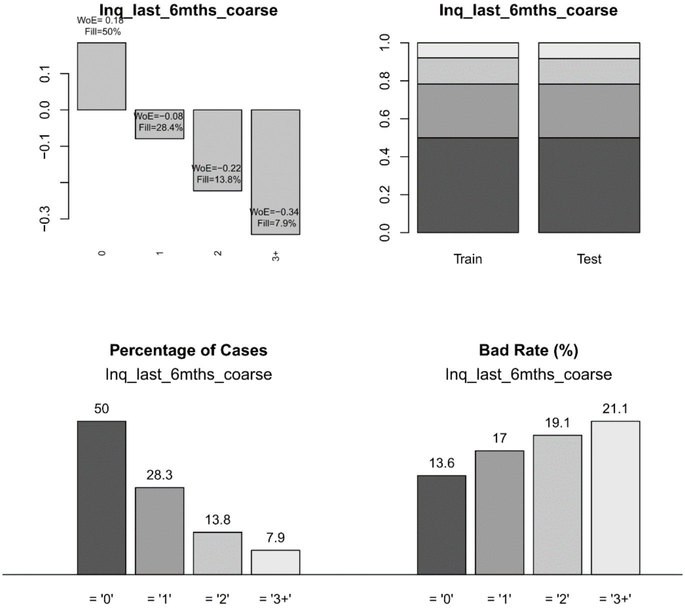
Source: Authors' calculations.

Figure 15: Variable *Term – coarse classing*



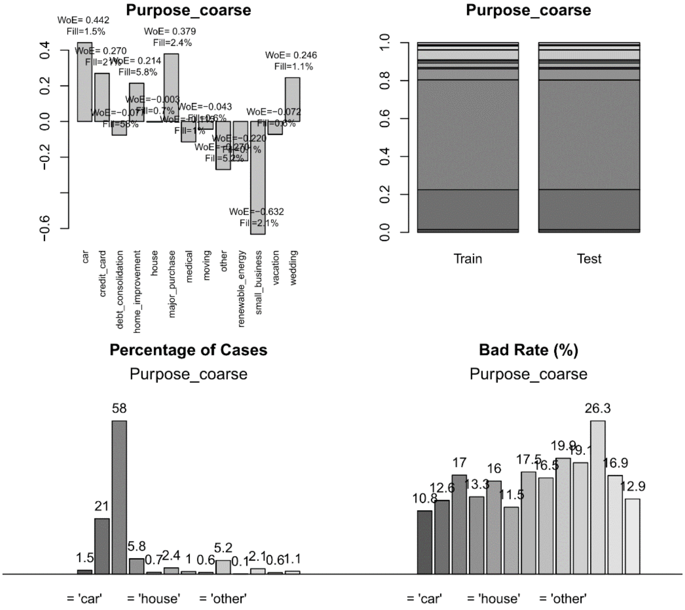
Source: Authors' calculations.

Figure 16: Variable *Inq_last_6mths – coarse classing*



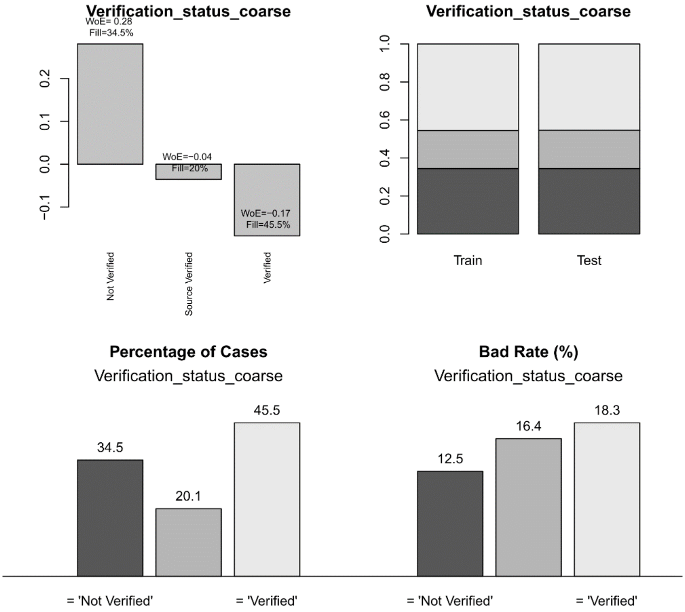
Source: Authors' calculations.

Figure 17: Variable *Purpose – coarse classing*



Source: Authors' calculations.

Figure 18: Variable *Verification_status – coarse classing*



Source: Authors' calculations.