Journal of Economics and Management Research. Vol. 9, 2020
https://doi.org/10.22364/jemr.9.02

DEVELOPMENT OF BORROWERS' SOLVENCY ASSESSMENT MODEL: LOGISTIC REGRESSION APPLICATION

DĀNIELS JUKNA
Mg. oec.

Abstract

Borrowers' solvency assessment models can not only increase company's profit, but also potentially decrease the impact from the negative economic consequences of the crisis. However, there is no consensus on such models. Considering the flaws in the scientific literature, the main aim of this article was to develop the borrowers' solvency assessment model, which can be applied in practice. The most appropriate method for developing such models was found to be logistic regression, and this research goal is to identify the best modelling approach to achieve the highest borrowers' solvency predictability. By implementing the best-chosen model, a nonbank lending company could provide a 42.5% lower total borrowers risk of default than without implementing such a model. Depending on the risk policy of the non-bank lending company, three methodologies were developed based on different assumptions about the significance of type 1 error and type 2 error in the company to determine the exact cut-off value.

Keywords: Latvia, lending, logistic regression, nonbank loans, solvency assessment

JEL classification: C31, C55, E51

INTRODUCTION

Issuing a loan is an unsafe business, but at the same time, it is one of the main activities and sources of income for financial institutions in Latvia. One of the biggest challenges for lenders is to develop effective lending and risk policies to ensure the desired return and to prevent a recurrence of the negative economic consequences of the 2007 US mortgage crisis, which included serious shortcomings in assessing borrowers' solvency.

Studies show that statistical models improve the accuracy of lending decisions and make lending more cost-effective, but there is no consensus on how to develop such models. Different publications use different methods, including different factors and select different model valuation indicators. In addition, many studies do not include or establish factor relative importance and do not conclude which factors are associated with good or bad borrower solvency. The shortcomings and incomplete information in the publications were considered to achieve

the highest possible results for the borrower solvency assessment model. By doing that answers to three questions are sought: (1) what factors allow to assess the borrowers' solvency; (2) which factor values are associated with a borrower with good solvency; (3) what are the relative importance of factors in the borrower solvency assessment model. Answers to these three questions provide statistical borrower solvency assessment model suitable for a non-bank lending company, which could be applicable in practice.

Research methods used in the study are statistical parametric (multifactor logistic regression) data processing method to create a borrower solvency assessment model, determining the relationship of independent variables with the borrowers' solvency and factor relative importance; mathematical programming method (GRG nonlinear) to determine the cut-off value of the best classes for the respective methodology.

A solvency assessment model for non-bank borrowers was developed using the logistic regression method and a step-by-step modelling approach with the Akaike information criterion. The obtained model can be applied in practice, considering the value of independent variables and certain cut-off value, which is obtained based on the risk policy of a non-bank lending company. The implementation of the obtained model would ensure a 14.6% low total risk of default of borrowers, which is 42.5% lower than without the introduction of such a model.

The article is structured as follows. Section 1 briefly reviews the literature on the topic of borrowers' solvency assessment. Section 2 presents the data and describes the methodology employed in this study. Section 3 provides the borrowers solvency assessment model results and efficiency of implementing it. The last section concludes.

LITERATURE REVIEW

Analysis of the borrower solvency assessment has received a fair share of attention in international literature (see for example Bolton (2009), Vidal and Barbon (2019)), however in case of Latvia the available literature is rather limited.

The solvency assessment of borrowers is based on a points system, which determines the risk level and solvency of a person, depending on the characteristics of the borrower, historical and current liabilities, and other factors (Anderson, 2007). The development of a borrower solvency assessment model can also be cost-effective, for example, one experiment with historical data in Bolivia concluded that rejecting 12% of loans disbursed to riskier customers in 2000 would have reduced the number of overdue loans by 28% (Schreiner, 2003).

The purpose of developing a risk score scale is to create a segmentation or index that can be used to classify consumers into two or more different groups (Bolton (2009) recommends to divide consumers into two groups – borrowers

with poor and good solvency), so econometric methods for modelling a dependent variable, as well as statistical classification methods are commonly used (Glennon *et al.*, 2008).

The borrowers' solvency models in the market are sufficiently general and standardised (Glennon *et al.*, 2008) and in Latvia they are not sufficiently explained, therefore making decisions based on these models is not recommended.

There may be a lag in the borrowers' solvency assessment model since this model does not include data for the whole population (applications rejected by the lender are not considered). To carry out this study and to develop an appropriate model, it is assumed that all other potential borrowers have similar characteristics as borrowers mentioned in this data set have. This assumption is based on the findings of the study that there were minimal differences in the accuracy of the model classification compared to the model based on the financial institution's existing customers and the model based on all applications (Banasik *et al.,* 2003).

Choosing the right factors is an important step in developing a solvency assessment model, as they will be used to predict the probability to default. It can be concluded that most of the publications mention the relationships of the variable that are not based on the results of the model.

Factors are used to distinguish borrowers with poor solvency can be distinguished from borrowers with good solvency by five categories of data: (1) demographic and educational indicators; (2) financial indicators; (3) employment indicators; (4) loan repayment discipline indicators; (5) other data (Mpofu & Mukosera, 2012).

The following relationships between demographic and educational indicators and borrower solvency can be found in the studies available in the literature: women are often found to be less risky men; the risk of default decreases with age and is also lower for married borrowers with dependents (possibly due to existing dual income (Schreiner, 2003)); homeowners represent a category of less risky borrowers because the house can be used as collateral; education is a very strong predictor of default, as borrowers with a higher level of education show much less default than other borrowers (Vojtek and Kocenda (2008)).

The financial indicators, for example, the amount of income, are important factors in assessing borrowers' solvency and are economically interpretable, as they show that borrowers with larger amounts of money have a lower risk of default (borrowers have more money to repay loan payments). However, another study found that risk of default does not depend on the absolute amount of income (i.e. the difference between income and liabilities), but it does depend on relative income (i.e. the ratio of expenses to income) (Vidal and Barbon (2019)). This means that high-income and high-spending borrowers are also risky.

The following relationships between employment indicators and borrower solvency can be found in the studies: self-employed people are often rated lower than employees (employment stability indicates payment stability), (Vojtek and

Kocenda (2006)); frequent changes in low-skilled jobs are riskier. On the other hand, some of the employment indicators, such as the employment sector or the number of years worked in an enterprise, are not statistically significant factors (Vojtek and Kocenda (2008)).

The following relationships between loan repayment discipline indicators and borrower solvency can be found in the studies: collateral reduces risk of default; recent loans are much riskier than those whose customers have a longer relationship with the financial institution (Vojtek and Kocenda (2006)); the risk of default increases with number of days overdue, delay status, debt to payment ratio and debt amount (Bolton, 2009). This information significantly reduces the problem of asymmetric information between the borrower and the lender.

In addition, studies provide information on the relationship between such other factors and the borrowers' solvency: the risk of default is higher if the borrower has applied at night (00:00–05:59), used the Android operating system; the risk of default is lower if the borrower has applied in the evening (18:00–23:59), used the IOS operating system or went to the lender's website through a direct channel; a borrower who uses a paid email provider is associated with good solvency; if the borrower has more than two or no telephone numbers, they have a higher risk of default than those who have only one telephone number; if the borrowers' e-mail consists of borrowers' personal information (for example, name, surname), then these borrowers have a lower risk of default (Berg *et al.*, 2019); loans for home improvement and renovation work are riskier than those for real estate (Vojtek and Kocenda (2008)).

DATA AND METHODOLOGY

The methods commonly used to assess solvency are econometric modelling or statistical methods. Historically, discriminant analysis and linear regression were the most widely used methods for determining solvency and are still used in the development of solvency assessment models in some studies as complementary methods, however, in recent years, logistic regression is probably the most widely used method for assessing solvency (Bolton, 2009).

The logistic regression model is an extension of the linear discriminant analysis, which allows overcoming problems with data abnormality. The logistic regression method can also deal with category data – the solution is to use dummy variables for each data category. One of the disadvantages of logistic regression is that this method is sensitive between explanatory variables, so it is necessary to make sure that there are no such variables in the training data set. Another disadvantage of this method is the sensitivity to missing values (all observations with missing values should be deleted) (Vojtek and Kocenda (2006)).

The risk of default is usually constructed in the form of an index so that all borrowers can be divided into two or more classes. Borrowers' solvency dependent value is usually dichotomous, consisting of two values – good solvency

and poor solvency (Samreen *et al.*, 2013) and it complies with the rules of logistic regression method (Bolton, 2009).

When developing borrowers' solvency assessment models, a threshold of 90 days is usually used to distinguish customers with good solvency from customers with bad solvency (Choy and Laik (2011)), but companies can also use a threshold of 30, 60 or even 180 days, two or three delayed payments or any other event that is related to a fall in a company's profitability. Defining borrowers with poor solvency is important before developing a solvency assessment model, as it directly affects the results of the model (Vidal and Barbon (2019)).

All available variables may be included in the borrower solvency assessment model, but various problems may be identified, such as multicollinearity between variables, the variables in the model are not statistically significant, the variables are not stable, and the variable has too few observations. These problems can be identified and remedied gradually by developing new models and deciding which one is the best, but there are ways in which these problems can be eliminated before borrowers' solvency assessment models are developed: (1) if an independent variable does not reach 1,500 borrowers with poor solvency, it is recommended to exclude it from the model; (2) factor analysis can be used to reduce the number of variables to be included in the borrower solvency assessment model and to avoid multicollinearity between variables (Samreen *et al.*, 2013); (3) information values is used to select important independent variables in binary models, ranking variables based on their significance (Bhalla, 2015).

It is considered that there is no optimal number of variables that should be used in the development of borrower solvency assessment models. Loan applications can include between 3 and 20 variables and in some cases the number is even higher (Berg *et al.*, 2019). To achieve sufficient predictability of a borrowers' solvency, the model should include at least 1500 borrowers with poor solvency, but the model may include fewer observations (e.g. 50–100) if no other observations are available. In this case the model should be reviewed regularly as information for new borrowers becomes available (Vidal and Barbon (2019)).

Factor analysis is method that could help to avoid multicollinearity between variables. However, it has drawbacks, for example, (1) this method does not allow to determine which exact variable from the obtained group of factors should be selected and used for modelling; (2) the factor analysis may show a high correlation between two important factors, such as the borrowers' income and the borrowers' credit liabilities, but excluding either of these factors, significant discrimination against borrowers may disappear, (3) the factor analysis also does not provide information on how much multicollinearity is already in the model (Samreen *et al.*, 2013). To avoid multicollinearity between variables without using factor analysis, the Variation-Inflation Factor (VIF) can be used. The VIF value in models should not exceed 10, however, it is also not recommended to use a model with this value that exceed 4 (Bock, 2019).

As the logistic regression model is one of the most used statistical techniques for solving the problem of binary classification, and for such models, independent variables can be estimated using information values using the formula:

$$IV = \sum (a - b) \times ln\left(\frac{a}{b}\right) \quad [1],$$

where *IV – information value,*
a – the ratio of properly classified borrowers with poor solvency,
b – the ratio of properly classified borrowers with good solvency.

The higher the value of the information, the greater the ability of the relevant independent variable to distinguish good solvency borrowers from poor solvency borrowers. When choosing independent variables to be included in the borrower solvency assessment model, it is recommended that the information value be higher than 0.10, which could indicate that the variable has moderate discrimination (predictive power), but the information values are sensitive enough to how the independent variable is grouped (Bolton, 2009). The threshold for including independent variables in the model could be chosen even lower, for example 0.02, which indicates that the independent variable has at least some predictive power with respect to the dependent variable (Tan, 2020). The reason for such a low threshold in one of the studies is that logistic regression model could include more available variables, for instance, social and demographic variables, despite the fact that they tend to present lower information values (Vojtek and Kocenda (2008)). One of the main disadvantages of information values is that they are assessed for each independent variable separately (Bhalla, 2015), which in turn means that the ability of the interaction of independent variables to distinguish borrowers with good solvency from those with poor solvency is not assessed.

Table 1

Information values for independent variables

Independent variable	Information value	Predictability
Education Level	0.5202	Very high
Email contains personal information	0.4940	High
The part of the day of filling in the application	0.2056	Moderate
Age	0.1238	Moderate
Purpose of the loan	0.0869	Poor
Email domain	0.0857	Poor
Industry	0.0777	Poor
Marital status	0.0712	Poor
Basic income	0.0690	Poor
Total income	0.0443	Poor

Independent variable	Information value	Predictability
Monthly credit liabilities	0.0381	Poor
Gender	0.0309	Poor
Application completion time (hours)	0.0278	Poor
Amount of outstanding debt	0.0129	No predictability
Additional income	0.0100	No predictability
Amount of outstanding debts for the last two years	0.0070	No predictability
Number of dependents	0.0019	No predictability

Source: author's calculations based on a non-bank lending company customers data

Using information criteria such as the Akaike Information Criterion (AIC), a gradual selection of the best models can be made. The selection of the variable in the final model is mainly based on successive statistical tests, and this approach is reliable and widely used for borrowers' solvency assessment models (Votjek and Kocenda (2008)). A lower AIC value indicates that the model is better given the number of independent variables and the number of observations. The comparison of AIC in the logistic regression model is based on relative increases rather than absolute values, but it is not determined what the optimal relative increase is to consider that there are no significant differences between the models (Date, 2019). To compare different models using AIC, the absolute difference formula of the respective model AIC and the minimum AIC found in the models is recommended (see formula 2). If this difference is greater than 10, then it indicates that there is a significant difference between the models (Burnham and Anderson (2004)).

$$\Delta_i = AIC_i - AIC_{min} \quad [2],$$

where Δ_i *– Akaike absolute difference of information criteria,*
AIC_i *– the Akaike information criterion for the respective model,*
AIC_{min} *– the minimum Akaike information criterion found in the models.*

Various indicators can be used to determine the suitability and predictability of borrowers' solvency assessment models, such as the Gini coefficient (the value should exceed 0.7) (Bolton, 2009), Pearson's Chi-squared, the Hosmer-Lemeshow test (the null hypothesis of the test is that the model is correctly specified) (Bartlett, 2014), the first and second error estimates (classification table with a specific cut-off value), the divergence statistical indicator, Tjur's coefficient of discrimination (Blochlinger and Leippold (2006)). Several indicators are compared to evaluate the models and the best model is adopted by analysing diagnostic tests and lender's risk policy. The best model can be chosen with a slightly higher Gini coefficient, but with a lower first-type error, as it may be more important for the lender to reduce the amount of loans issued to borrowers

who default. This means that the lender's risk policy is decisive in choosing the best models.

Unpublished customers data from a non-bank lending company were used in the study. For the development of the borrower solvency assessment model, a total of 20570 borrowers to whom the Non-Bank Lending Company has issued loans between January 2017 and December 2019 are available. 70% of all available data (14399 observations) is used to develop the borrower solvency model, while 6171 observations are used to assess the predictability and discrimination of the model (this sample is not used during the model development phase). The data set is randomly divided into model development and test data sets using R studio software. In the model development data set, borrowers with poor solvency is about 25% (10842 borrowers with good solvency and 3557 borrowers with poor solvency).

MAIN FINDINGS

In this study, several logistic regression models have been developed and the best ones have been selected for which predictability, stability and suitability have been tested. These models provide a basis for testing an important hypothesis that could help to understand the best modelling approach (see table 2).

Table 2

Summary of all models and their basic indicators

Model number	Description	Sign. level for variables	AIC	Max. VIF
1.	Model with the largest number of independent variables	At least 95%	11063	1.511
2.	Model without independent variable *monthly credit liabilities*	At least 95%	11065	1.511
3.	A model which is developed based on information values	At least 95%	11151	1.512
4.	A model which is developed based on step-by-step AIC-based approach and information values	At least 95%	11066	1.513
5.	Model with the smallest number of independent variables and insignificant differences in AIC	At least 95%	11073	1.513

Source: author's calculations based on a non-bank lending company customers data

The first model has been chosen as one of the models to be used in the further analysis, as it includes the most independent variables and their classes (dummy variables) compared to the other models, and all their coefficients have reached a 95% significance level (this model is considered to be a comparable model because it also has the lowest AIC).

In the second model the independent variable *monthly credit liabilities* has been removed. This model satisfies the AIC difference indicator, as it does not exceed 10. The amount of the borrowers' credit liabilities is usually not known exactly at the time of loan issuance, because not all lenders provide information on the borrower' s credit liabilities to all credit information offices in Latvia (Credit information bureau, 2020; Consumer Rights Protection Centre, 2020). This model could be used to determine whether, without knowing the borrowers' credit obligations, the borrowers' solvency could be predicted just as well.

The information values of the independent variables in the literature are considered as one of the main methods indicating which independent variables are important to predict borrowers' solvency. The third model showed a significant difference between the Akaike information criterion comparing to the first model, but this does not yet indicate that the model based on information values is less discriminatory. To make sure that a step-by-step modelling approach is better than modelling initially from information values, the third model is chosen as one of those for which the model's suitability and predictability are tested.

Given the information values, it has been found that *the purpose of the loan for home improvement* is an important factor in assessing the solvency of borrowers, but the first and second models do not include this factor. If only step-by-step modelling AIC-based approach were used, borrowers' solvency assessment models would not include the above-mentioned factor. To check whether the information values have added value in the development of models (by adding the independent variable *home improvement*), the fourth model is also chosen in the further analysis. For this model, the AIC difference does not exceed 10, so there is no significant difference between the first and fourth models.

As the fourth model did not show a significant difference in the Akaike information criterion between the first model, a fifth model was developed in which two independent variables are excluded from the fourth model. In order to find out which of the two models (the one with the largest number of independent variables or the one with the smaller number of independent variables) provides better results and solvency predictability, the fifth model is also chosen in the further analysis. It should be noted that the AIC difference for this model does not exceed 10, so there is no significant difference between the first and fifth models.

The factor signs of all models that have chosen for further analysis are summarised in Table 3. All the obtained models show the same signs for the independent variables, which indicates that the independent variables are sufficiently stable and the inclusion or exclusion of different independent variables from the models does not affect the relationship of the independent variable to the dependent variable or borrowers' solvency. It also shows that there is no multicollinearity in the models (this is indicated by VIF that does not exceed 5 in any model).

Table 3

Relationships of independent variables with dependent variables in different models

Independent variable / model	1.	2.	3.	4.	5.
Age	+	+	+	+	+
Man	–	–	–	–	–
Basic education	–	–	–	–	–
Vocational education	–	–	–	–	–
Higher Education	+	+	+	+	+
Married	+	+	+	+	+
Living together	–	–	–	–	–
Email domain – Inbox.lv	–	–	–	–	–
Email contains personal information	+	+	+	+	+
Vehicle repair	–	–			
Purchase of consumer goods	–	–			
Home improvement			+	+	+
Health Care	–	–		–	–
Refinancing	–	–	–	–	–
Art, recreation, entertainment	–	–		–	–
Education	+	+		+	+
Extraction and processing of materials	–	–	–	–	–
Trade	–	–		–	–
The application is completed at night	–	–	–	–	–
The application is completed in the evening	–	–		–	–
Application completion time	+	+	+	+	+
Basic income	+	+	+	+	+
Monthly credit liabilities	+		+	+	
Outstanding debt	–	–		–	–
Debts paid	+	+		+	

Source: author's calculations based on a non-bank lending company customers data

To determine which of the models has the highest discrimination, predictability, and which of them can be considered the best model, all five models are initially compared using six criteria: sum of squared deviations, Pearson's Chi-squared (this corresponds to Pearson's Chi-squared test), Hosmer-Lemeshow test, Gini coefficient (corresponding to the Kolmogorov-Smirnov test), Tjur's coefficient of discrimination and percent of cases correctly classified (hereafter – classification indicator) for the test population (cut-off = 0.5).

The evaluation indicators of the models are presented in Table 4 and calculated using R studio software. These indicators are similar for all models and this indicates that relatively small changes have been made to the models, which does

not significantly affect the suitability and predictability of the models. The sum of squared deviations and Pearson's Chi-squared do not provide any statistical interpretation but ranking them can lead to a conclusion as to which model is better. The Hosmer-Lemeshow test for all five models exceeds 5%, which shows that the H0 hypothesis that there is no bad model for predicting a dependent variable cannot be rejected, so all five models are suitable for forecasting. The Gini coefficient is high enough for all models (it exceeds 70%, which is the recommended value for the model to be predictive). Also, Gini coefficient is not so high (less than 90%) that there may be some errors in the data. The Tjur's coefficient of discrimination is used to estimate the coefficient of determination, which is about 35% in all models (except for the third model, where this coefficient is about 34%). This shows that in the model the dependent variable is explained on average around 34%–35% of the independent variables. In addition, the classification indicator of the test data set has been assessed, which has been calculated by dividing the correctly classified borrowers by the total number of borrowers. The cut-off value for good and bad solvency is chosen to be 0.5, but this does not mean that this value is determined as the best. By choosing a cut-off value of 0.5, all models have achieved a classification indicator about 83%, which indicates that in 83% of cases the borrowers in the test set would be classified correctly (borrowers' solvency would be correctly determined), while in 13% the model would give an erroneous result.

Table 4

Comparison of evaluation indicators of five models

Indicator	Model 1	Model 2	Model 3	Model 4	Model 5
Sum of squared deviations	11012.6	11017.4	11116.6	11017.7	11028.6
Pearson's Chi-squared	14384.0	14376.4	14453.7	14304.4	14351.2
Hosmer-Lemeshow test	65.08%	64.95%	64.67%	64.95%	64.84%
Gini coefficient	85.21%	85.22%	84.99%	85.19%	85.18%
Tjur's coefficient of discrimination	35.09%	35.06%	34.38%	35.00%	34.95%
Classification indicator (cut-off = 0.5)	82.95%	82.94%	82.63%	82.82%	82.87%

Source: author's calculations based on a non-bank lending company customers data

To be able to compare all five models with each other, these models are ranked according to the above six model evaluation indicators. A score of "1" indicates that the model has the highest score in the respective indicator, while a rank of "5" indicates that the model has achieved the lowest score in the respective indicator. The values of the ranking evaluation indicators of five models are presented in Table 5. Considering ranked evaluation indicators, the best model is the first model, while the worst model is the third model.

Table 5

Comparison of ranked evaluation indicators of five models

Indicator	Model 1	Model 2	Model 3	Model 4	Model 5
Sum of squared deviations	1	2	5	3	4
Pearson's Chi-squared	4	3	5	1	2
Hosmer-Lemeshow test	1	3	5	2	4
Gini coefficient	2	1	5	3	4
Tjur's coefficient of discrimination	1	2	5	3	4
Classification indicator (cut-off = 0.5)	1	2	5	4	3
Average	1.67	2.17	5.00	2.67	3.50
Rank by 5 indicators	1	2	5	3	4

Source: author's calculations based on a non-bank lending company customers data

The first model and the second model have very similar results, and they are among the two best models. This suggests that the exclusion of the borrowers' monthly credit liabilities from the model did not lead to such a significant deterioration. If databases that cost enough are used to determine the monthly credit liabilities, then the Non-Bank Lending Company may choose to assess the solvency of the borrowers using the second model without obtaining information from the databases on the borrowers' monthly credit liabilities.

The third model, compared to other models, has the lowest results in all model evaluation indicators, so this model has the lowest discriminability. This suggests that a step-by-step modelling approach is better than the initial modelling from information values. It is valuable to use information values to identify comparable variable models or to comment possible correlations, but it is not possible to rely entirely on them and build a model based on information values, as this gives lower results compared to other approaches.

The fourth model additionally includes the purpose of the loan for home improvement, which according to the information values can predict the dependent variable, but this model has worse results than the first and second models. This means that information values have not provided added value in the development of models.

The fourth model has a higher predictability than the fifth model in terms of model evaluation indicators, which shows that the number of independent variables in the respective models is not excessive and that the addition of each new statistically significant independent variable provides better predictability (this could be due to large amount of data and adding new independent variable does not significantly reduce the number of degrees of freedom and thus does not lead to model redundancies).

The final model (the first model) shows a high rate of discrimination of the dependent variables, as Gini coefficient is 85.21%. The values and statistics of independent variables for this model are shown in Table 6. Considering the independent variables of the model, solvency can be explained by 35.09% (it is estimated using the Tjur's coefficient of discrimination). If the cut-off is chosen to be 0.5 for the final model, then the classification indicator of the tested data is estimated at 82.95% (the indicator is obtained from the classification tables).

Table 6

Values and statistics of independent variables of the final model

Variables	Value	Standard error	Z statistics	P-value	Sign. level
Intercept	1.1520	0.1212	9.505	0.0000	***
Age	0.0236	0.0022	10.537	0.0000	***
Male	-1.1570	0.0488	-23.707	0.0000	***
Basic education	-2.2320	0.0749	-29.8	0.0000	***
Vocational education	-0.9960	0.0652	-15.272	0.0000	***
Higher education	0.3282	0.0663	4.954	0.0000	***
Married	0.3215	0.0617	5.208	0.0000	***
Living together	-0.2049	0.0583	-3.516	0.0004	***
Email domain - Inbox.lv	-0.5773	0.0547	-10.557	0.0000	***
Email contains personal information	1.3890	0.0497	27.945	0.0000	***
Vehicle repair	-0.2567	0.0767	-3.349	0.0008	***
Purchase of consumer goods	-0.3007	0.0678	-4.433	0.0000	***
Health Care	-0.5948	0.0707	-8.419	0.0000	***
Refinancing	-0.8360	0.0781	-10.703	0.0000	***
Art, recreation, entertainment	-0.2830	0.1163	-2.432	0.0150	*
Education	0.4092	0.1652	2.477	0.0132	*
Extraction and processing of materials	-0.3705	0.0619	-5.982	0.0000	***
Trade	-0.2443	0.0675	-3.618	0.0003	***
The application is completed at night	-2.7240	0.1562	-17.443	0.0000	***
The application is completed in the evening	-0.2464	0.0558	-4.414	0.0000	***
Application completion time	0.0073	0.0015	4.91	0.0000	***
Basic income	0.0003	0.0001	3.837	0.0001	***
Monthly credit liabilities	0.0004	0.0002	2.193	0.0283	*
Outstanding debt	-0.0005	0.0001	-4.469	0.0000	***
Debts paid	0.0000	0.0000	2.132	0.0330	*

"***" – more than 99.9% significance, "**" – more than 99% significance,
"*" – more than 95% significance, "." – more than 90% significance, " " – less than 90% significance.

Source: author's calculations based on a non-bank lending company customers data

For the model to be applicable to a non-bank lending company, it is necessary to find out what is the most appropriate cut-off value to separate borrowers with good solvency from borrowers with poor solvency. As the risk policy for a non-bank lending company does not specify which criteria must be considered to select a most appropriate cut-off, three different methodologies are proposed in the work:

1. minimising the first type of error (minimise the potential risk of lending to borrowers with poor solvency, thus increasing the overall loan repayment rates);
2. minimising the second type of error (minimise the unearned profit from those borrowers who have good solvency, but who were rejected, thus issuing loans to as many borrowers as possible with less risk); and
3. minimising the first and second types of error (maximise classification indicator).

As shown in Table 7 for all methodologies has achieved its set goal. The first methodology has the smallest first type of error (4.88%) compared to other methodologies, while with this methodology company would issue 3609 loans (58.5% of applications), which is significantly less compared to other methodologies, which would provide at least 1,400 more loans. The second methodology has the smallest second type of error (0.94%) compared to other methodologies, thus using this methodology company would issue a loan almost after every application (93.5% of applications). It should be noted that this methodology will lead to low repayment rates (the repayment rate will be even lower than the second type of error, as this model only estimates borrowers who have become customers, but did not take into account rejected applications). The third methodology achieves the highest classification indicator, which means that if 0.508 is selected for the cut-off, then this is the scenario in which the solvency of the borrowers is determined most accurately. The first type of error has been significantly reduced in the model compared to actual non-performing borrowers (from 25% to 12%), which means that the model is able to discriminate against borrowers well enough to reduce the number of loans to non-performing borrowers by about twice.

Table 7

Obtained error, classification indicators for three methodologies

Methodology	Cut-off	Error type 1	Error type 2	Classification indicator	Issued loans
Methodology 1	0.800	4.88%	20.99%	74.14%	3609
Methodology 2	0.201	19.82%	0.94%	79.24%	5768
Methodology 3	0.508	11.89%	5.10%	83.00%	5022

Source: author's calculations based on a non-bank lending company customers data

If the non-bank lending company has not developed a risk policy and it is not determined which type of error is more significant, then it is recommended to use the third methodology model, in which the first and second types of error are balanced. A classification table has been developed for this methodology (see Table 8).

Table 8

Classification table for the third methodology

Prognosis	Actual	
	0	**1**
0	834	315
1	734	4288

Source: author's calculations based on a non-bank lending company customers data

Using the logistic regression method in the third methodology, the model incorrectly classifies 1,049 borrowers, of which 734 borrowers are the ones to whom the loan will be issued, but they would have poor solvency. The default risk in this case would be 14.62%, which is lower than if this model were not implemented.

Most of the variables included in the final model are often used in theoretically based models of borrower behaviour and solvency assessment in the literature, as well as the economic significance of the estimated coefficients has been assessed in the analyses performed in this work and based on the literature. The final borrowers' solvency assessment model can be applied in practice, considering the value of independent variables and a specific cut-off value, which is obtained based on the risk policy of the non-bank lending company.

CONCLUSION

The logistic regression is the most appropriate method in practice as in borrower solvency models the dependent variable is binary. To avoid subjective lending decisions and to make the results of the model statistically significant, easy to interpret and be traceable over time, it is necessary to define a specific set of rules that classify borrowers with poor solvency.

Solvency assessment models (1) may include all available independent variables and progressively sought issues that may affect the model or (2) may not include these issues in the first model using a variety of approaches and methods. With a gradual modification of the models, higher discriminability can be achieved than with the elimination of problems in the first model.

Several indicators are compared to evaluate different models, and the best model is adopted based on the analysis of diagnostic tests and based on the lender's risk policy, however, the lender's risk policy is decisive in the selection of the best models.

The borrowers' monthly credit liabilities could be excluded from the model without significantly reducing predictability and discriminability. Such models are important to the lender because the amount of the borrowers' credit obligations is usually difficult to determine and change over time.

A step-by-step modelling approach is better than initially building models from information values. The values of the information are valuable to use to identify comparable variable models or to comment on possible correlations in the models.

Information values did not have added value to the development of models, as models that were created without the involvement and influence of information values showed higher results. This repudiate that information value is one of the best methods for selecting independent variables in binary models.

The number of independent variables in the respective models is not excessive and the addition of each new statistically significant independent variable provides better predictability (in general, a large amount of data is available on borrowers, which does not reduce the number of degrees of freedom relatively).

Using six different indicators of model suitability, predictability, and discriminability, the best model is the one that uses a step-by-step approach estimating the Akaike information criterion with as many statistically significant independent variables as possible.

Independent variables may indicate good or poor solvency, but the existence of a single independent variable indicating poorer solvency does not immediately indicate to the borrower that the borrower is poor solvency, as the model is designed with all factors in mind.

For the model to be applicable to a non-bank lending company, it is necessary to find out what is the most appropriate cut-off value. If the non-bank lending company does not have a detailed risk policy, it is recommended to use a methodology in which the first and second types of errors are equally important and in which the highest classification indicator is achieved.

The obtained borrower solvency assessment model can be applied in practice, considering the value of independent variables and certain cut-off value, which is obtained based on the risk policy of a non-bank lending company. The implementation of the obtained model would ensure a 14.6% low total risk of default of borrowers, which is 42.5% lower than without the introduction of such a model.

REFERENCES

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

Banasik, J., Crook, J., Thomas, L., 2003. *Sample Selection Bias in Credit Scoring Models.* Journal of the Operational Research Society, 54 (3), 822–832 p.

Bartlett, J., 2014. *The Hosmer-Lemeshow goodness of fit test for logistic regression.* [online] Available at: <https://thestatsgeek.com/2014/02/16/the-hosmer-lemeshow-goodness-of-fit-test-for-logistic-regression/> [Accessed 11 March 2020].

Berg, T., Burg, V., Pu, M., Gombovic, A., 2019. *On the Rise of FinTech's – Credit Scoring using Digital Footprints*. The Review of Financial Studies, hhz099, 66 p.

Bhalla, D., 2015. *Weight of Evidence (WOE) and Information Value (IV) Explained.* [online] Available at: <https://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html> (Accessed 25 February 2020).

Blochlinger, A., Leippold, M., 2006. *Economic Benefit of Powerful Credit Scoring.* Journal of Banking & Finance, 30 (3), 851–873 p.

Bock, T., 2019. *What are Variance Inflation Factors (VIFs).* [online] Available at: <https://www.displayr.com/variance-inflation-factors-vifs/> (Accessed 25 February 2020).

Bolton, C., 2009. *Logistic regression and its application in credit scoring.* Pretoria: University of Pretoria.

Burnham, K. P., Anderson, D. R., 2004. *Multimodel Inference – Understanding AIC and BIC in Model Selection.* Sociological Methods & Research, 33 (2), 261–304 p.

Choy, M., Laik, M. N., 2011. *A Markov Chain approach to determine the optimal performance period and bad definition for credit scorecard.* International Journal's Research Journal of Social Science and Management, 1 (6), 227–234 p.

Consumer Rights Protection Centre, 2020. *Companies that have received a special permit (license) for the provision of consumer credit services.* [online] Available at: <http://www.ptac.gov.lv/lv/table/kapitalsabiedribas-kuras-sanemu-licenci-pateretaju-krediteanas-pakalpojumu-sniegsanai> (Accessed 12 January 2020).

Credit information bureau, 2020. *Sources of Credit Information.* [online] Available at: <http://www.kib.lv/individuals/sources-of-credit-information/> (Accessed 15 January 2020).

Date, S., 2019. *The Akaike Information Criterion.* [online] Available at: <https://towardsdatascience.com/the-akaike-information-criterion-c20c8fd832f2> (Accessed 16 January 2020).

Glennon, D., Kiefer, N. M., Larson, E. C., 2008. *Development and Validation of Credit-Scoring Models.* Journal of Credit Risk, 4 (3), 70 p.

Mpofu, T. P., Mukosera, M., 2012. *Credit Scoring Techniques: A Survey*. International Journal of Science and Research, 3 (8), 165–168 p.

Pritchard, J., 2018. *How Credit Scores Work and What They Say About You.* [online] Available at: <https://www.thebalance.com/how-credit-scores-work-315541> (Accessed 12 December 2019).

Samreen, A., Sarwar, A., Zaidi, F. B., 2013. *Design and Development of Credit Scoring Model for the Commercial Banks in Pakistan: Forecasting Creditworthiness of Corporate Borrowers.* International Journal of Business and Commerce, 3 (17), 1–26 p.

Schreiner, M., 2003. *Scoring: The Next Breakthrough in Microcredit.* Washington: Consultative Group to Assist the Poor/World Bank Group.

Vidal, M. F., Barbon, F., 2019. *Technical Guide – Credit Scoring in financial inclusion.* Washington: Consultative Group to Assist the Poor/World bank.

Tan, J., 2020. *Weight of Evidence (WoE) and Information Value (IV)*. [online] Available at: <https://towardsdatascience.com/model-or-do-you-mean-weight-of-evidence-woe-and-information-value-iv-331499f6fc2> (Accessed 25 February 2020).

Unpublished materials of a non-bank lending company.

Vojtek, M., Kocenda, E., 2006. *Credit Scoring Methods.* Czech Journal of Economics and Finance, 56 (3–4), 152–167 p.

Vojtek, M., Kocenda, E., 2008. *Default Predictors in Retail Credit Scoring: Evidence from Czech Banking Data.* Emerging Markets Finance and Trade, 47 (6), 80–98 p.