Application of the Logistic Regression in Non-bank Financial Institutions

M. Papoušková, V. Pacáková

***Abstract*—**The aim of this paper is to demonstrate the application of the logistic regression in non-bank financial institutions. Compared to banks, these companies have a significant share of badly paying clients. The dependent variable is so called default. Among independent variables which are included into the model are variables like net income, installment, age, available funds, marital status and more. The methods of logistic regression assess the intensity of the impact of independent variables on the ability to repay the loan. This information is essential for managing of the credit risk of non-banking institutions.

***Keywords*—**Credit Risk Management, Logistic Regression, Non-bank Financial Institutions, Risk

I. Introduction

INDEBTEDNESS is a phenomenon of our society. In this modern time, the debt is perceived as a resource to accelerate the achievement of the objectives in an effort to increase their standard of living. Many people tend to finance their credit needs. There are a large number of credit products as well as there are many financial institutions and agents who provide credit. People with poor credit ratings are rejected in banking institutions and come to a non-bank financial institution that provide their loan products with great alacrity and often does not look at the financial options for the future of the debtor. Higher risk of these applicants means a higher price of credit. Applications that come to apply for loans to non-bank financial institutions often have low financial literacy, do not estimate their financial options, do not pay attention to the terms of the contract and often do not understand them.

It is necessary to deal with the "responsible lending", meaning that creditors should properly assess the creditworthiness of the client and to properly evaluate the suitability of the product.

Valuation the creditworthiness of applicants is key. Underestimation of the creditworthiness of an applicant means an increase in the probability of default. Default is defined by Basel II like: "A default is considered to have occurred with regard to a particular obligor when either or both of the two following events have taken place:

M. Papoušková is with the Institute of Mathematics and Quantitative Methods, University of Pardubice, Czech Republic (e-mail: monika.papouskova@upce.cz).

V. Pacáková is with the Institute of Mathematics and Quantitative Methods, University of Pardubice, Czech Republic (e-mail: viera.pacakova@upce.cz).

• The bank considers that the obligor is unlikely to pay its credit obligations to the banking group in full, without recourse by the bank to actions such as realizing security (if held).

• The obligor is past due more than 90 days on any material credit obligation to the banking group. Overdrafts will be considered as being past due once the customer has breached an advised limit or been advised of a limit smaller than current outstanding."

Credit risk is the risk that arises from the inability or unwillingness to meet its obligations, and should be the dominant interest of creditors.

Credit scoring is the process of assigning a single quantitative measure, or score, to a potential borrower representing an estimate of the borrower's future loan performance [3]. Credit scoring is a process that is used for prediction default.

One of the most frequently used method for the development of credit scoring models in the market is logistic regression.

II. Logistic Regression

Logistic regression is one method of multivariate statistical analysis. The approach of this technique is based on the concept that each single attribute should be tested before inclusion in the model. Logistic regression is possible to divide into several categories: binomial logistic regression, multinomial logistic regression and ordinal logistic regression.

The outcome of binomial logistic regression is binary or dichotomous. On the other side multinomial logistic regression is used to predict outcomes with more than two categories.

The dependent variable is binary, nominal or ordinal variable and independent variables are categorical or continuous variable.

The general formula is as follows:

$$Logit(p_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \quad (1)$$

where:

p is probability of default,

$x_1, \ldots x_k$ are independent variables,

β_0 is constant,

$\beta_1, \ldots \beta_k$ are coefficients,

Logit is $ln(p/(1-p))$.

More information about logistic regression can be found in

[4], [5]. For the first time logistic regression for predicting whether a client falls into default or not was used in [6]. Father use of logistic regression in credit scoring was in [1], [9].

III. Data

Application of the binary logistic regression is presented for data file that contains 19 839 records. The data file includes information about provided consumer loan of non-bank financial institutions during the period September 2012 and June 2014. Each variable was categorized. Recoding categorical explanatory variables is necessary because the label of categories have to make sense in interpreting the results [6]. The dependent variable in this case is variable called Default. In the table 1 can be seen used independent variables.

Table 1: Independent variables

Region of residence	1 – 14, 51
Installment (currency = CZK / 1 month)	0 – 12 respectively: <=500, 501-1000, 1501-2000, 2001-2500, 25001-3000, 3001-3500, 3501-4000, 4001-4500, 4501-5000, 5001-5500, 5501-6000, 6000+
Amount disbursed (CZK)	0 – 10 respectively: <=10000, 10001-20000, 20001-30000, 30001-40000, 40001-50000, 50001-60000, 60001-70000, 70001-80000, 80001-90000, 90001-100000, 100000+
CNCB (group)	0 – 9 respectively: 12, 13, 14, 16, 17, 25, 166-320, 321-412, 413-495, 496+
Age	18-83
Available funds (CZK / 1 month)	0 – 6 respectively: <=0, 1-4000, 4001-8000, 8001-12000, 12001-16000, 16001-20000, 20001+
Maturity (months)	0 – 6 respectively: 12, 18, 24, 30, 36, 42, 48
Marital status	0 – 4 respectively: Single, Married, Divorced, Widow/er, Other
Net income (CZK / 1 month)	0-290000
Number of flags in SOLUS	1 – 5 respectively: 0, 1, 2, 3, 4
Typ of client	0-2 respectively: Retirement, Parental/Maternity leave, Employee
Guarantor	Yes, No
SOLUS (separately A, D, B, P, C, U, Z)	Yes, No

Source: own

Association SOLUS is the register which contains negative information about the clients who are not interested in paying their contractual liabilities and/or about the clients who have trouble with paying their debts to a member company of the Association SOLUS.

CNCB (Czech Non-Banking Credit Bureau, z.s.p.p.) is the register that includes negative and positive information about the clients.

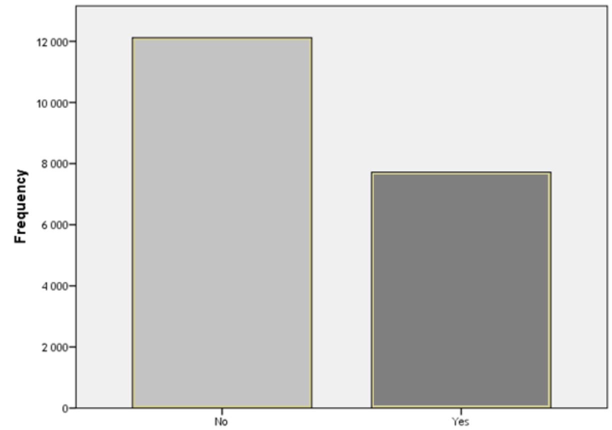

Figure 1: Distribution of client according to default/non-default

In the table 2 or the figure 1 is possible to see the distribution of client according to default/non-default.

Table 2: Distribution of client according to default/non-default

		Freq	Percent	Valid Percent	Cumulative Percent
Valid	No	12122	61,1	61,1	61,1
	Yes	7720	38,9	38,9	100,0
	Total	19842	100,0	100,0	

Source: own (in IBM SPSS Statistics)

IV. Problem Solution

In this paper is used binary logistic regression, specifically method Forward Selection (Likelihood Ratio), i.e. stepwise selection method with entry testing based on the significance of the score statistic, and removal testing based on the probability of a likelihood-ratio statistic based on the maximum partial likelihood estimates in IBM SPSS Statistics software.

After 14 steps, the classification table looks as follows:

Table 3: Classification Table

Observed		Predicted		
		Default		Percentage Correct
		No	Yes	
Default	No	10 164	1 955	83,9
	Yes	4 896	2 824	36,6
Overall Percentage				65,5

Source: own (in IBM SPSS Statistics)

Besides the constant, variables include into the model are net income, available funds, installment, maturity, age, SOLUS A, SOLUS B, SOLUS D, SOLUS P, SOLUS U. region of residence, marital status, CNCB.

Their values are seen in the table 4.

Table 4: Variables in the Equation

	B	S.E.	Wald	Sig.
Age	**-,019**	**,002**	**111,136**	**,000**
Available funds			**48,461**	**,000**
Available funds (1)	-,063	,322	,039	,844
Available funds (2)	-,203	,323	,393	,531
Available funds (3)	-,330	,325	1,028	,311
Available funds (4)	-,451	,329	1,880	,170
Available funds (5)	-,588	,335	3,079	,079
Available funds (6)	-,765	,342	5,008	,025
Installment			**361,164**	**,000**
Installment (1)	,070	1,135	,004	,951
Installment (2)	,255	1,134	,050	,822
Installment (3)	,485	1,134	,183	,669
Installment (4)	,574	1,134	,256	,613
Installment (5)	,734	1,135	,418	,518
Installment (6)	,844	1,135	,553	,457
Installment (7)	,938	1,135	,684	,408
Installment (8)	,993	1,135	,765	,382
Installment (9)	,969	1,137	,727	,394
Installment (10)	1,209	1,137	1,131	,288
Installment (11)	1,303	1,144	1,298	,255
Installment (12)	1,455	1,136	1,640	,200
Maturity			**79,097**	**,000**
Maturity (1)	,594	,359	2,742	,098
Maturity (2)	,703	,357	3,879	,049
Maturity (3)	,825	,358	5,319	,021
Maturity (4)	,928	,356	6,784	,009
Maturity (5)	,956	,362	6,996	,008
Maturity (6)	1,097	,358	9,361	,002
Client type			**6,664**	**,036**
Client type (1)	,125	,128	,954	,329
Client type (2)	-,101	,050	4,175	,041
SolusA	**1,010**	**,110**	**84,240**	**,000**
SolusB	**,517**	**,045**	**134,596**	**,000**
SolusD	**,444**	**,069**	**41,502**	**,000**
SolusP	**,322**	**,130**	**6,142**	**,013**
SolusU	**,254**	**,130**	**3,842**	**,050**
Residence client			**54,595**	**,000**
Residence (1)	,127	,088	2,074	,150
Residence (2)	,383	,119	10,388	,001
Residence (3)	-,132	,097	1,848	,174
Residence (4)	,139	,099	1,952	,162
Residence (5)	,060	,086	,474	,491
Residence (6)	,041	,096	,184	,668
Residence (7)	-,176	,098	3,194	,074
Residence (8)	,020	,095	,046	,830
Residence (9)	,028	,097	,083	,774
Residence (10)	,035	,087	,162	,688
Residence (11)	,180	,092	3,842	,050
Residence (12)	,174	,104	2,797	,094
Residence (13)	,172	,108	2,561	,110
Residence (14)	1,018	,453	5,043	,025
Marital status			**54,673**	**,000**
Marital status (1)	-,232	,045	26,454	,000
Marital status (2)	,039	,050	,610	,435
Marital status (3)	-,033	,078	,181	,670
Marital status (4)	,253	,448	,318	,573
CNCB			**387,646**	**,000**
CNCB (1)	-,273	,045	37,194	,000
CNCB (2)	-,512	,048	115,945	,000
CNCB (3)	-,685	,065	111,114	,000
CNCB (4)	,880	1,324	,442	,506
CNCB (5)	-,457	,221	4,284	,038
CNCB (6)	,359	,076	22,003	,000
CNCB (7)	,245	,051	23,467	,000
CNCB (8)	1,040	,696	2,230	,135
CNCB (9)	,193	,078	6,133	,013
Net income	**,000**	**,000**	**6,196**	**,013**
Constant	**-,648**	**1,236**	**,275**	**,600**

The column labeled like B in the table 4 presents the values of the regression coefficients of the logistic model. Using these coefficients the probability of default is possible to estimate for any combination of values of the explanatory variables. The column Sig. informs that the explanatory variable significantly influences the probability of default. The important information provides *exp(B)*. For example, if age is increased by one unit, then the odds of default is 0,982 times higher, i.e. *exp(-0,019)*.

Next example of interpretation is: If marital status is divorced then odds of default is 1,04 times higher than odds of default of single clients or if client type is parental/maternity leave then odds of default is 1,133 times higher than odds of default of pensioners in this data set.

In the figure 2 is possible to see the ROC curve. Area under the ROC curve is 0,679.

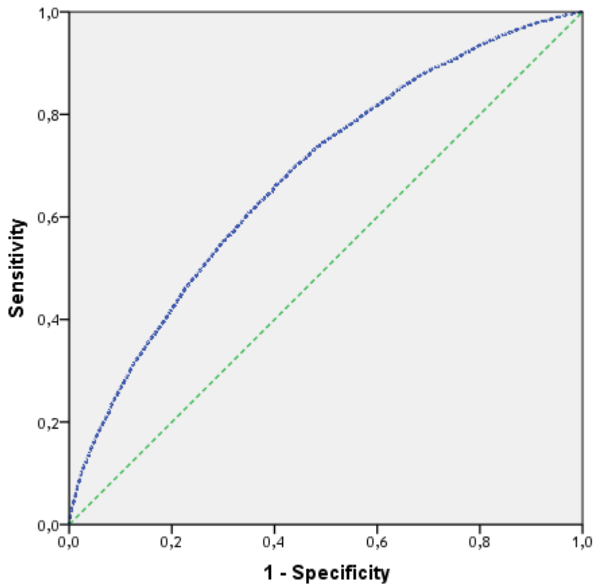

Figure 2: ROC Curve

V. CONCLUSION

During the construction of credit scoring model (score cards) there are some issues that have to be sorted out. Current problems are how to define the default, select the input sample, respectively from when to when to choose loans for modeling, what to do in case of missing or outliers.

The quality of the clients of non-banking financial institutions begins where the quality of the clients in the banks ends. This means that the number of bad clients is much higher.

Therefore, risk management in these institutions is more difficult in comparison with the banks and need to use the various quantitative methods for the assessment of creditworthiness of the clients. One of the frequently used methods is currently logistic regression. The article provides application of this method to real data from the Czech Republic and the application of its results.

REFERENCES

[1] Berkowitz J., Hynes R. Bankruptcy Exemptions and the Market for Mortgage Loans. The Journal of Law and Economics 42.2., 1999, pp. 809-830.
[2] Dzelihodzic, A., Donko D. Data Mining Techniques for Credit Risk Assessment Task. Recent Advances in Computer Science and Applications,
[3] Feldman, R. Small business loans, small banks and big change in technology called credit scoring. The Region Sep. 1997, pp. 19-25.
[4] Harrell, Frank E. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. Springer Science & Business Media, 2013.
[5] Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. Applied logistic regression. Vol. 398. John Wiley & Sons, 2013.
[6] Orgler, Yair E. A credit scoring model for commercial loans. Journal of Money, Credit and Banking 2.4., 1970, pp. 435-445.
[7] Pacáková, V.; Stankovičová, I.; Pacák, B. Application of Logistic Regression in Financial Sector. Proceedings of the 12th Polish-Slovak-Ukrainian Scientific Seminar organized by the Department of Demography of the Cracow University of Economics (November 8-10, 2005, Krynica): A Comparative Analysis of the Socio-economic Consequences of Transition Processes in Central and Eastern European Countries, Krakow: AE, 2006.
[8] Sarlija, Natasa, Mirta Bensic, and Marijana Zekic-Susac. "Modeling customer revolving credit scoring using logistic regression, survival analysis and neural networks." Proceedings of the 7th WSEAS international conference on neural networks. World Scientific and Engineering Academy and Society (WSEAS), 2006.
[9] Steenackers, A., and M. J. Goovaerts. A credit scoring model for personal loans. Insurance: Mathematics and Economics 8.1., 1989, pp. 31-34.