Risk Assessment Technology of Crediting with the Use of Logistic Regression Model

Rostyslav Yurynets[1][0000-0003-3231-8059], Zoryna Yurynets[2][0000-0001-9027-2349], Dmytro Dosyn[3][0000-0003-4040-4467], Yaroslav Kis[4][0000-0003-3421-2725]

[1] Lviv Polytechnic National University, rostyslav.v.yurynets@lpnu.ua
[2] Ivan Franko National University of Lviv, zoryna_yur@ukr.net
[3] Lviv Polytechnic National University, dmytro.dosyn@gmail.com
[4] Lviv Polytechnic National University, yaroslav.p.kis@lpnu.ua

Abstract. The mathematical model of loan borrower estimation taking into account expert assessment has been created. The application of the model based on logistic regression aimed at estimating the solvency of the client has allowed to establish the correlation between risk factors and probable size of loan risk.

Keywords: machine learning, credit risk, scoring model.

1 Introduction

Loan scoring which provides flexible tools of estimation of loan risks as well as a possibility to automatize the adoption of decisions connected with loans is one of the main approaches towards the quantitative evaluation of the solvency of borrowers which has been actively developing worldwide. Information systems of credit scoring based on the comparative data analysis of credit history of the existing borrowers and of the applicants for a loan analogically, allow to define integrated score assessment of solvency (reliability) of potential borrowers. The purpose of loan scoring is optimization of a decision making concerned with granting the bank loans which in its turn causes current relevance of our research.

Practical functioning of risk management systems at Ukrainian banks is characterized by a number of imperfections which added many problems to banking institutions during crisis and partially caused its emergence. Thus, the methods of the analysis and assessment of risk of borrowers on behalf of ownerships were imperfect. As a result many consumer loans were issued to persons who could not cope to return the them.

Currently, the potential of tools for scoring technologies of risk assessment related to crediting has not been properly used. Appropriate use of scoring technologies throughout all life cycle of the loan allows to make more adequate and justified decisions which can be efficiently automated.

Nowadays the application of special information technologies which would enable the chance at a stage of credit applications consideration directed at the eliminating of unreliable borrowers is relevant for the bank sphere. The Ukrainian banks use mainly a

single application scoring. This situation is explained by both underdeveloped scoring systems in Ukraine and high prices for services of scoring model developers.

The world practice uses different scoring types at different stages of crediting in addition to the stage of consideration of the loan application, continuing to estimate the properness of the loan application during all lifecycle of a loan, and even at a stage of collecting debts and transfers to collectors. In particular, there is an analytical module for the formation of models of solvency assessment of borrowers and their segmentation as well as the possibility of algorithm development and fraud identification models in a complex control system of loan scratches of SAS Credit Scoring for Banking. The following indicates provided possibilities of realization of all types of scoring for different types of tasks and for different types of data.

Loan scoring includes the models of decision making and the main methods which provide support to creditors during the process of decision making about a consumer loan. The use of these methods allows to define the potential loan owner, the size and expeditious strategy that will aid to increase profitability of borrowers as well as to estimate a possible risk. Credit scoring is based on real data that allows to refer it to reliable estimates of personal solvency.

The scoring occurs in several types depending on the tasks which have to be solved: Application scoring (scoring of the applicant) is the assessment of client solvency to receive the loan (scoring according to biographical data is in priority); Behavioral scoring (scoring of behaviour) is used for the assessment of return reliability of the issued loans (the behavioural analysis); Collection scoring (scoring aimed to analyze arrears) is used to assess a possibility of the complete or partial return of the loan in case of violation of debt repayment periods (calculation of risks behind portfolio contents); Fraud scoring (scoring aimed to fight swindlers) is used to assess the probability of a new client to be a swindler; Response scoring (scoring of a response) is the assessment of consumer's reaction (response) to the directed offer; Attrition scoring (scoring of losses) is the assessment of further product implicational probability or a product supplier change.

Traditionally, two approaches are used in case of realization of a scoring system in Ukraine. The first one is classical (retrospective) scoring on the basis of the analysis of historical data with application of the modern mathematical methods if such analysis enables the choice of the significant fields for the questionnaire designed for the borrower and other indexes. The second approach is an expert scoring when, for instance, the expert sets solvency estimation rules and the software automatizes this algorithm without application of any statistical methods for the analysis of historical data. Nowadays the second alternative is the most common in both medium-sized and small banks and in many larger ones. Nevertheless, in recent years the first option has been characterized by the market demand due to the appearance of small but considerably more important number of credit stories in some markets.

Currently the integration of several scoring methods is the most efficient, namely statistical and expert. The fact is that a system includes and realizes the tools which give the chance to integrate the statistical approach and expert scoring, to consider regional specifics of the market and loan products as well as to talk about their effective use.

The modern set of methods of loan scoring is developed on the basis of the tools for the predictive analysis which belongs to a wide range of so-called methods of the profound data analysis (data mining).

The tools for the predictive analysis include:

- the statistical methods designed on the basis of discriminant analysis (the linear regression, logistic regression);
- different options of the linear programming;
- classification tree;
- neural networks;
- genetic algorithm;
- method of the closest neighbors.

2 Main Material Presenting

The problem of model operation, assessment and management of credit risk was investigated by the following Ukrainian and foreign scientists: G.I. Beregova, R. Gallati, V.M. Gorbachuk, A.B. Kaminsky, B.Yu. Kishakevich, A.A. Lobanov, O. Habyyuk, A.V. Chugunov and other.

Model operation of loan risk is a significant and currently important problem, since the application of effective loan risk assessment model allows the financial organization to save time and money and to protect themselves from undesirable losses or even a default. In addition, it helps to adopt management decisions concerning the avoidance or minimization of the negative influence caused by the tendency to risk. Therefore, the problem related to the choice of an optimal assessment model for loan as well as the application of new loan assessment and modelling methods remains unresolved problem risk in the modern changeable environment.

The article aims at developing and realizing of mathematical evaluation model of the loan borrower taking into account expert assessment.

The evaluation of client solvency by the model of logistic regression in the presented paper required the statement of correlation between risk factors and probability size of loan risk y which has values ranging from 0 to 1. The selection of the most informative quantitative variables of financial risk included the possible usage of the discriminant analysis tools. The Table 1 presents a matrix fragment of values of nine indexes which were selected during the analysis in which a binary variable y describes the following situations: 0 – overdue or problematic loan; 1 – in due time repaid loan.

The tab. 1 introduces designations: d – average monthly income, s – the sum of the loan, t – the term of the loan, r – an interest rate, B – the age, O – expert's assessment of a professional, economic and social status of the client.

The calculation of solvency coefficient of the ownership is performed by a formula [8]:

$$x_1 = \frac{d}{s\left(\frac{1}{t}+\frac{r}{12\cdot 100}\right)} \quad (1)$$

Table 1. Matrix fragment of values indexes for the evaluation of client solvency

No.	*y*	*d*, UAH	*s*, UAH	*t*, months	*r*	B, years	O
9	1	13200	50000	24	11	37	5
10	0	12410	50000	12	10	57	3
11	1	10718	25000	24	13	49	4
12	1	9623	44170	72	14	30	4
13	0	11000	60000	36	12	47	3
14	0	15000	70000	12	10	65	4
15	1	8109	29360	36	12	29	4
16	1	9720	48760	24	12	34	5

We use the prepared data which are partially presented in tab. 2 for the estimation of unknown parameters of econometric model [1]:

$$P(y_i = 1|x_i) = F(b_0 + b_1x_{i1} + b_2x_{i2} + b_3x_{i3}) + \varepsilon_i,\ i = 1,2,\ldots,n \qquad (2)$$

in which $P(y_i = 1 \mid x_i)$ is the probability that the i-th value of a binary variable is 1 under the condition of $\boldsymbol{x}_i$; $F(z) = \frac{1}{1+e^{-z}}$ – logistic function; ε_i – random component; x_1 – individual creditworthiness ratio; x_2 – age; x_3 – expert assessment of the profession and the socio-economic status of the client.

Logistic function has that property that its values are ranging from zero to one at any values of an argument.

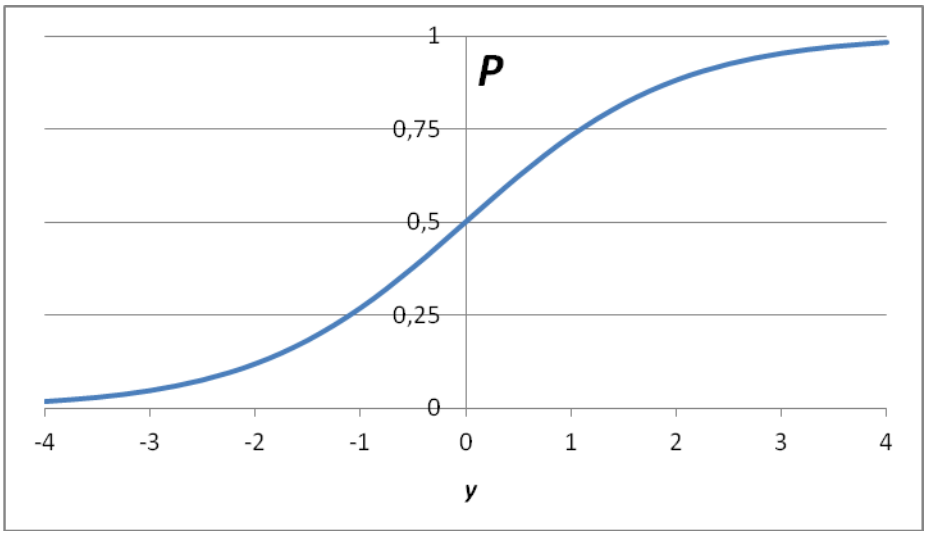

Fig. 1. Logistic curve

Table 2. Matrix fragment of the value indicators

y	x_1	x_2	x_3
1	5,19	37	5
0	2,71	57	3
1	8,17	49	4
1	8,53	30	4
0	4,85	47	3
0	2,34	61	4
1	7,31	29	4
1	3,86	34	5

Suppose that there is a training selection received in test data: (x_{i1}, x_{i2}, x_{i3}, y_i). It is required to estimate the parameters in equation (2) using the obtained sample. For the estimation of unknown parameters use of the maximum likelihood principle. According to this principle, the values that give a maximum for the likelihood function are taken as parameter estimates. The likelihood function in our case is as follows:

$$L(\boldsymbol{y}, \boldsymbol{b}) = \prod_{y_i=1}^{n} F(\boldsymbol{x_i b})^{y_i} \, [\mathbf{1} - F(\boldsymbol{x_i b})]^{1-y_i} \tag{3}$$

Here the following designations are for short accepted:

$$\boldsymbol{b} = (b_0, b_1, b_3)^T, \tag{4}$$

$$\boldsymbol{x}_i = (1, X_{i1}, X_{i2}, X_{i3}), \tag{5}$$

$$\boldsymbol{x_i b} = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} \tag{6}$$

Usually, instead of function (3), its logarithm is used, which does not change the essence of the problem, but allows one to get rid of the multiplication:

$$\ln L = \sum_{i=1}^{n} \boldsymbol{y_i} \ln F(\boldsymbol{x_i b}) + (\mathbf{1} - \boldsymbol{y_i}) \ln(\mathbf{1} - F(\boldsymbol{x_i b})), \tag{7}$$

where n is the number of tests.

Obviously, function (7) has a maximum. For estimation of values of the parameters at which the maximum of the function (7) is reached, the partial derivatives are calculated from these parameters and equated to zero:

$$\frac{\partial ln L(\boldsymbol{y}, \boldsymbol{b})}{\partial \boldsymbol{b}} = 0 \tag{8}$$

Thus, the task has been reduced to solving the system of equations (8) with respect to the unknown parameters b. The solution to this system presents certain difficulties, since it is not linear.

For the solution, you can use the Newton – Raphson method. The method involves the selection of some initial approximation of the solution and its consistent improvement in the course of performing a series of calculations:

$$\boldsymbol{b}_{t+1} = \boldsymbol{b}_t - \frac{\partial \ln L(\boldsymbol{b}_t)}{\partial \boldsymbol{b}} \left[\frac{\partial^2 \ln L(\boldsymbol{b}_t)}{\partial \boldsymbol{b}\, \partial \boldsymbol{b}'}\right]^{-1} \tag{9}$$

where

$$\frac{\partial \ln L(\boldsymbol{b})}{\partial \boldsymbol{b}} = \left(f_0(\boldsymbol{b}),\, f_1(\boldsymbol{b}), \ldots,\, f_m(\boldsymbol{b})\right) \tag{10}$$

$$f_0(\boldsymbol{b}) = \sum_{i=1}^{n} F(\boldsymbol{x_i b}) - \sum_{\{i:Y_i=1\}} 1 \tag{11}$$

$$f_j(\boldsymbol{b}) = \sum_{\boldsymbol{i=1}}^{n} F(\boldsymbol{x_i b}) x_{ij} - \sum_{\{i:Y_i=1\}}^{n} x_{ij}\,, j \;=\; 1{,}2{,}3 \tag{12}$$

$$\frac{\partial^2 \ln L(\boldsymbol{b}_t)}{\partial \boldsymbol{b}\, \partial \boldsymbol{b}'} = \tag{13}$$

$$= \begin{pmatrix} \sum_{i=1}^{n} F(\boldsymbol{x_i b})(1 - F(\boldsymbol{x_i b}))\,, & \sum_{i=1}^{n} F(\boldsymbol{X_i b})(1 - F(\boldsymbol{x_i b}))x_{i1}, & \ldots & \sum_{i=1}^{n} F(\boldsymbol{x_i b})(1 - F(\boldsymbol{x_i b}))x_{i3}, \\ \sum_{i=1}^{n} F(\boldsymbol{x_i b})(1 - F(\boldsymbol{x_i b}))x_{i1}, & \sum_{i=1}^{n} F(\boldsymbol{X_i b})(1 - F(\boldsymbol{x_i b}))x_{i1}x_{i1}, & \ldots & \sum_{i=1}^{n} F(\boldsymbol{x_i b})(1 - F(\boldsymbol{x_i b}))x_{i3}x_{i1}, \\ \ldots\ \ldots\ \ldots & \ldots & & \ldots \\ \sum_{i=1}^{n} F(\boldsymbol{x_i b})(1 - F(\boldsymbol{x_i b}))x_{i3}, & \sum_{i=1}^{n} F(\boldsymbol{x_i b})(1 - F(\boldsymbol{x_i b}))x_{i1}x_{i3}, & \ldots & \sum_{i=1}^{n} F(\boldsymbol{x_i b})(1 - F(\boldsymbol{x_i b}))x_{i3}x_{i3} \end{pmatrix}$$

The initial values can be defined as a vector of linear regression parameters:

$$\boldsymbol{b}_0 = (\boldsymbol{x}^T \boldsymbol{x})^{-1} \boldsymbol{x}^T \boldsymbol{y} \tag{14}$$

The computer preparation of data which includes calculation of the indexes which were selected during the discriminant analysis was carried out in the environment of the Excel application and further imported to a STATISTICA application for the analysis in a submodule "*Logistic regression*" of the Nonlinear Estimation module.

Parameters of the received model on the basis of logistic regression in the environment of STATISTICA have the following appearance: Model is: logistic regression (logit). Dependent variable: y. Independent variables: 3. Loss function is: maximum likelihood Final value: ,019218181.

-2*log(Likelihood): for this model=,0384364; intercept only=22,18071;

Chi-square = 22,14227; df = 3; p = ,0000611/

Apparently from parameters, the three-factor model provides high reliability which is confirmed by a calculated value a chi-square (22.14) and almost zero reliability not to reject a null hypothesis. Analytical expression of the constructed model will have such expression:

$$P(y_i = 1|x_i) = (1 + e^{-1{,}31 - 0{,}55x_{i1} + 1{,}31x_{i2} - 15{,}95x_{i3}})^{-1} \tag{15}$$

The adequacy of the designed model can be defined by the index of the McFadden likelihood ratio, using a formula

$$LRI = 1 - \frac{\ln L(\boldsymbol{b})}{\ln L(\boldsymbol{b_0})} = 0{,}97 \tag{16}$$

$\ln L(\boldsymbol{b})$ is a maximum value of a logarithmic function of credibility. It is reached in a point with coordinates equal to estimation parameters model $b = (b_0, b_1, b_2, b_3)$. $\ln L(\boldsymbol{b_0})$ is a value of a logarithmic function of credibility calculated by the assumption that $b_1 = b_2 = ... = b_m = 0$. The calculated value of the index of the McFadden likelihood ratio demonstrates adequacy of the designed model.

Using the constructed model, the possibility of granting loans to customers was calculated depending on the individual's solvency ratio for fixed values of the other variables (Fig. 2) ($y1$: $x_2 = 45$, $x_3 = 3{,}4$; $y2$: $x_2 = 55$, $x_3 = 4$; $y3$: $x_2 = 65$, $x_3 = 5$)

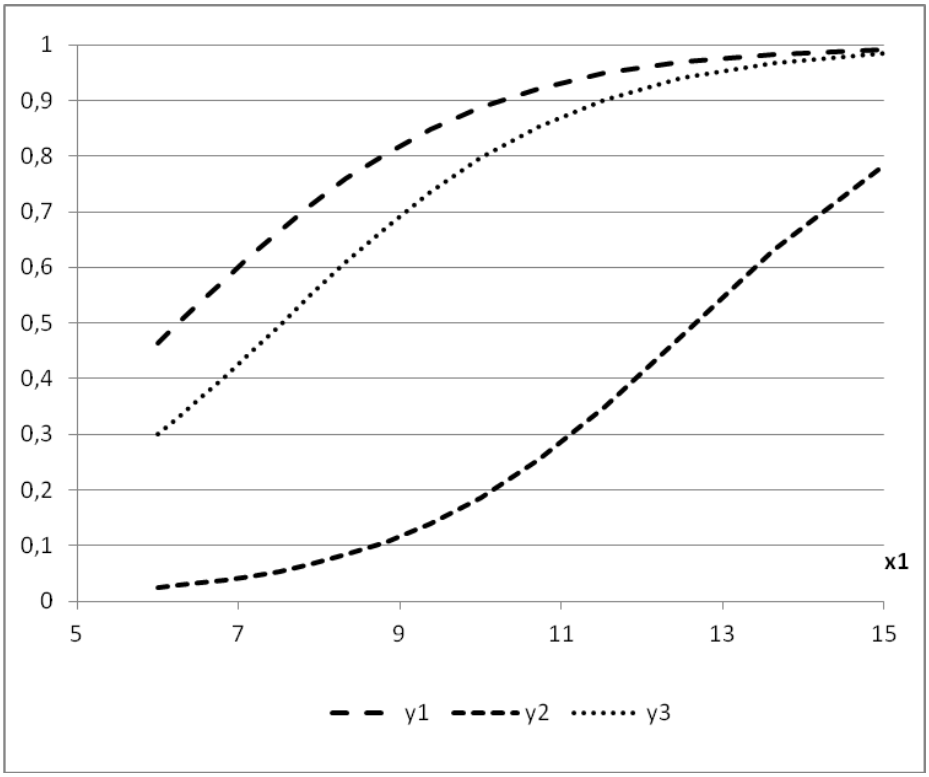

Fig. 2. Possibility of granting loans to customers depending on the individual's solvency ratio

The calculations of the possibility of issuing loans to customers, depending on the age of an individual with fixed values of the other variables (Fig. 3) ($y1$: $x_1 = 6$, $x_3 = 4$; $y2$: $x_1 = 15$, $x_3 = 3$; $y3$: $x_1 = 3$, $x_3 = 5$)

Using the constructed model, the possibility of granting loans to customers was calculated depending on expert assessment of the profession and the socio-economic status of the client for fixed values of the other variables (Fig. 2) (y1: x1 = 2, x2 = 25; y2: x1 = 2, x2 = 40; y3: x1 = 2, x2 = 55)

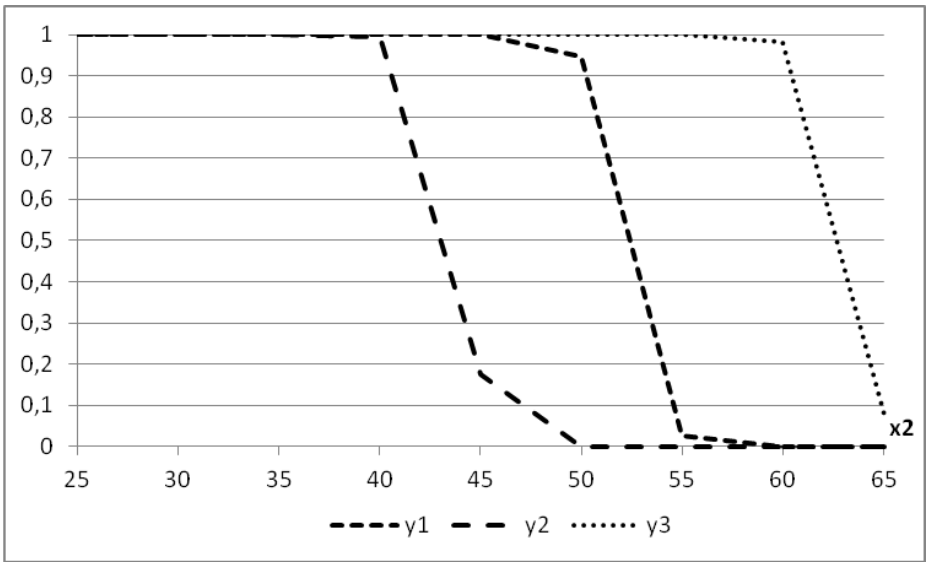

Fig. 3. Possibility of issuing loans to customers depending on the age of an individual

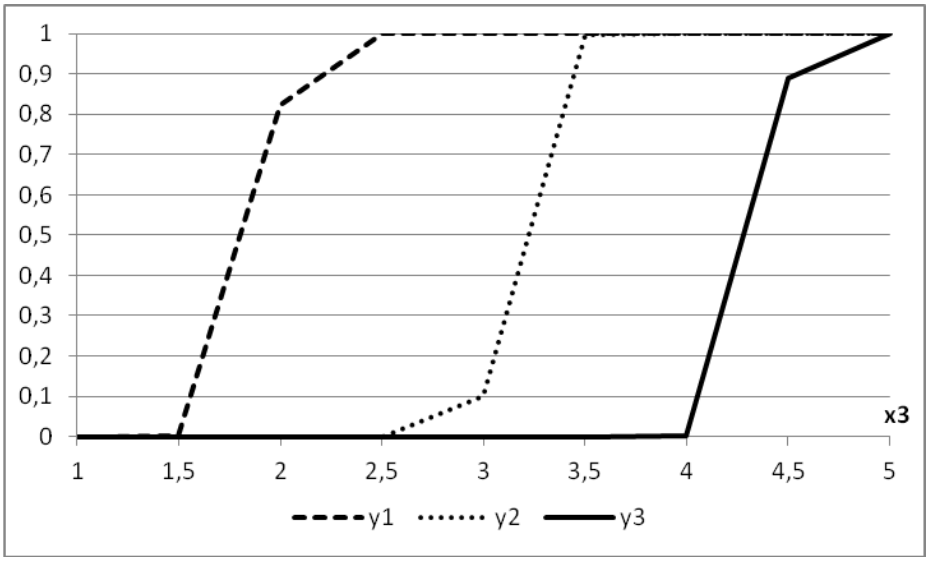

Fig. 4. Possibility of granting loans to customers depending on expert assessment of the profession and the socio-economic status of the client

The estimation of the possibility of insuring loans to new clients by means of the designed model, is provided in table 3.

Table 3. Value of indexes for assessing the solvency of new clients

№	d, UAH	s, UAH	t, months	r	B, years	O
1	11500	28000	18	11	59	4
2	9000	23000	12	10	35	3
3	14000	32000	24	12	52	4

$$P(y_1 = 1) = (1 + e^{-1{,}31-0{,}55\cdot 6{,}35+1{,}31\cdot 59-15{,}95\cdot 4})^{-1} = 0{,}00016, \quad (17)$$

$$P(y_2 = 1) = (1 + e^{-1{,}31-0{,}55\cdot 4{,}27+1{,}31\cdot 35-15{,}95\cdot 3})^{-1} = 0{,}996, \quad (18)$$

$$P(y_3 = 1) = (1 + e^{-1{,}31-0{,}55\cdot 8{,}47+1{,}31\cdot 52-15{,}95\cdot 4})^{-1} = 0{,}83. \quad (19)$$

The carried-out calculations indicate a possibility of granting the loan only to the second and third clients.

3 Conclusions

Loan risks carry the greatest danger to commercial banks in the context of providing and maintaining their financial stability. Therefore, introduction of new methods of assessment, management and consequently, as well as prevention of loan risks have to be the priority direction of development of a Ukrainian banking system.

The designed model of loan scoring allows the bank loan analyst to have an opportunity of justified and grounded self-decisions on loan service for customers and management of the loan portfolio under the conditions of intense competition in the market of retail loan issuing.

With the passage of time, any statistical model becomes inaccurate. It occurs for many reasons: owing to business cycles, changes of the customer data base of bank, structural shifts in economy, inflation etc.

Using the jargon of probability model it means that the influence of borrower's characteristics on the probability of returning or not returning the loan does not remain constant, and tends to change over time. Thus, in order to assure a continuous functioning of the scoring model, it requires periodical adjustment.

References

1. Greene, W. H.: Ekonometric analysis. Kyiv: Osnovy, 1196 p. (2005)
2. Altman, E. I., Saunders, A.: Credit risk measurement: Developments over the last 20 years. In: Journal of Banking and Finance, 21 (11-12), pp. 1721-1742 (1998)
3. Bashko, V. I.: Estimation of foreign exchange risks and cost of external state borrowing in Ukraine. In: Finance of Ukraine, 9, pp. 94-104 (2012)
4. Giesecke, K., Shilin, Z.: Transform analysis for point processes and applications in credit risk. In: Mathematical Finance, Vol. 00, 0, pp. 1-21 (2012)
5. Dolinsky, L. B.: Probabilistic models of default borrowers. In: Finance of Ukraine, 10, pp. 73-81 (2012)

6. Kaminsky, A. B.: Modeling of financial risks. Kyiv: Publishing and Printing Center "Kyiv University", 304 p. (2006)
7. Melnyk, O. H.: Methodical provisions on the rapid diagnosis of the threat of bankruptcy of the enterprise. In: Finance of Ukraine, 16, pp. 108-116 (2010)
8. Prokopenko, I.F., Ganin, V.I., Solyar, V.V., Maslov, S.I.: Fundamentals of Banking: Manual. Kyiv: TsNL, 410 p. (2005)