Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 2458-9403
Vol. 4 Issue 8, August - 2017

Forecasting The Risks Of Individuals Crediting Using Mathematical Models

Mykhailenko Viktor[1], Kozhukhivska Olha[2], Tetiana Prykhodko[3], Ivan Basiuk[3], Denys Nevinskyi[4]

[1] Kiev National University of Construction and Architecture, Kyiv, Ukraine
[2] The Bohdan Khmelnytsky National University of Cherkasy, Cherkasy, Ukraine
[3] National Aviation University, Kyiv, Ukraine
[4] Lviv Polytechnic National University, Lviv, Ukraine (*nevinskiy90@gmail.com*

***Abstract* — In this Paper, modern approaches to mathematical modeling of the retail crediting process are considered. As a result of processing actual statistical data regarding characteristics of bank clients has been established that acceptable forecasting results could be reached with Bayesian networks, binary nonlinear models and the internal rating based approach. All these methods have demonstrated high quality of classification the clients into two groups – those who return and those who will not return the credit. It has also been established the necessity of constructing a specialized decision support system on the system analysis principles to be used while crediting the potential clients. The advantages of such systems are possibility for effective preliminary data processing, the use of several alternative methods of clients' state estimation and the set of quality criteria at each step of data analysis. A high computing power of modern computers will provide a possibility for fast and objective client state estimation and make correct decisions.**

Keywords — retail crediting, mathematical models, client state estimation, decision making.

I. INTRODUCTION

One of basic directions of bank institutions activity is crediting of individuals. The proper management of such financial processes allows setting mutually advantageous long-term relations between the borrowers of credit and financial institutions. Considerable accumulations of bank capitals are quite up to date; expansions of banks' proposals and increase of organizational level and standardization of business processes promote the dynamic growth of amount of retail clients and volumes of credits. Also, there is simplification of banks requirements to the target audiences and reduction of time necessary for decision making to possibility of a person crediting. Obviously, the influence of such factors results in growth of losses as a result of the proper financial risks realization. If growth rates and revenue of credit brief-case are high enough, they fully recover financial losses as a result of realization of the risks. For this reason, most financial organizations during the protracted period did not do the proper investments in development of modern effective methods of control of crediting and introduction of modern information computer technologies, directed in decision making support in the process of management of the retail crediting risks. However, a case of retail crediting had been gradually changing to worse, the situation was aggravated by world financial crisis despite its (mainly) external origin.

From autumn 2008 most financial organizations faced the following problems: 1) – substantial limitations appeared in relation to access to the currency financial resources; absence of resources means limitation of crediting volumes, and consequently diminishing volumes and rates of growth of credit brief-case profitableness; 2) – as a result of crisis in the economic sphere (growth of currency rates, decline of level of labor payment, partial or complete loss of work, decline of production volumes) the volumes of problem credits grew considerably; 3) – the amount of swindle cases was increased in the process of crediting.

All these factors mean the necessity to change the existing approaches to organization and accompaniment of process of the retail crediting. Today there is an urgent need to create the effective principles of management and reliable (as for the results of calculations) computer information systems of making decisions support. The need to use modern methods of statistic and intellectual analysis of data has become especially demanded as well as mathematical modeling of financial and economical processes in order to build up mathematical models for predicting the possibility of credit returning.

On the whole, the crediting process consists of such stages [1–4]: 1) the evaluation of solvency of client; 2) the accompaniment and monitoring of process of payment of the taken credit; 3) the realization of measures in relation to the penalty of outstanding debt; 4) the analysis of current status of credit brief-case and making out the proper managing influences; 5) the permanent update (adaptation) of models of evaluation of client's solvency to the new terms.

Therefore, in this work there will be considered modern models and methods of evaluation of solvency of individuals on the basis of which there will be a possibility to build up the computer systems of making decisions support with the purpose of acceleration of data analysis processes and increase of objectivity and quality of decisions.

II. THE ANALYSIS OF EXISTING DEVELOPMENTS AND PROBLEM DEFINITION

Purpose of work is the following: to execute the analysis of modern methods of model's construction to evaluate the solvency of individuals; to choose modeling methods for the evaluation of solvency of the retail crediting clients on the basis of borrower's descriptions; to execute computing experiments in predicting client's solvency and compare the results.

The modern bank information systems already contain functions for credit requests treatment (application processing system – APS). Those are programmed complexes for making decision support in crediting. Actually, APS is a modern mean for analytical support of realization of credit processes with the use of plural of rules of credit decisions acceptance. Such programmed complexes are a certain constructor which introduction means the possibility of valuable organization of the process of credits delivery. However, this complex has certain limitations:

- functional limitations in the process of credits delivery: the system includes only those parameters and mechanisms which a developer considers sufficient for an effective management risks; if a customer needs new functions and mechanisms; their introduction is possible only by modification and addition of program code.
- relative difficulty in introduction of new processes and reengineering of existing processes; the process of introduction can last from four months to one year.

There exist programmed complexes which contain the certain solutions of mentioned tasks, for example, system Experian. Such complexes have limitations in creation of hierarchical (multilevel) strategies of borrowers' solvency evaluation. For example, with the help of such system it is a possibility to develop and implement operation procedures depending on credit history, however it is impossible to develop the mechanism of authentication of credit history on the basis of present payments information. It means that such complexes cannot be the unique module of decision making which can be used by an analyst to develop or realize the process of reengineering of difficult strategies of decision making.

Therefore, in the process of practical tasks there is a necessity of planning individual systems of making decisions support on the basis of plural of mathematical models which computer-integrated usage will provide effective high-quality support at making decisions in crediting of individuals.

III. MATHEMATICAL MODELS FOR CREDIT SCORING

Credit scoring means the analysis of client's solvency. The process of decision making in relation to the possibility of credit delivery is based on knowledge and information about clients. Modern data bases and knowledge bases of making decisions support information systems contain the row of determinations formulated by crediting experts and directed on explanation of information value, present models and possibilities of their use in a process of crediting. This information describes not only well-known possibilities of data analysis and requirements to crediting but also special knowledge (methods, models and calculation algorithms), their interpretation, internal terminology of financial institutions which concerns solving crediting issues.

On the whole, the models of credit scoring can be divided into two big categories: parametrical and non-parametrical. The group of parametrical includes: 1) linear probabilistic models; 2) models of binary choice; 3) models based on discriminant analysis; 4) neural networks; 5) neural illegible models; 6) buyer's networks. Non-parametrical scoring models include: 1) – models which are used in solving tasks of mathematical programming; 2) – classification trees (recursive classification) algorithm; 3) – models which are used in realization of method of the nearest neighbor; 4) – analytical hierarchical process of decision making; 5) - indistinct logic (and indistinct logic in combination with other procedures of decision making); 6) – expert estimation and systems on its basis.

Linear probabilistic models. Linear probabilistic model (LPM) is a model in a form of linear regression which dependent variable has the meaning from 0 to 1 depending what decisions are made concerning issuance of credit. Formally such model is represented in the following way:

$$y(k) = b_1 x_1 + b_2 x_2 + \ldots + b_m x_m + \varepsilon(k) \quad (1)$$

y – deepen dent variable the meaning of which corresponds to the decision making; $x_i, i = 1,\ldots,m$ - explaining variables (the characteristics of a client); b_i – coefficients (parameters) of regression equation which are estimated by data characterizing clients; $\varepsilon(k)$ – accidental process caused by existence of uncalculated disturbance and also mistakes of structure estimation and model parameters; k – client's identifier. In vector form equation has a following form:

$$y = b^T x + \varepsilon \quad (2)$$

Thus, conditional probability of credit receiving can be written the following way:

$$\Pr(y \mid x) = b^T x \quad (3)$$

This conditional probability can be interpreted as a probability to receive credit on condition of x information. Thus, after equation of calculation parameters (1) the latter can be used to estimate the probability of getting credit to a new client. The received estimation can be further compared with boundary value in order to make final decision concerning issuance of credit. LPM usage has such shortages: 1) – the variable possibility of getting meanings outside of

intervals [0, 1]; 2) – simultaneous usage in the right part category variables and variables which are represented by real numbers can lead to displacement of model estimates parameters; 3) – the process of crediting is more often characterized by linear dependents which needs the usage of models of other structures. Obviously, those shortages can cause getting rough estimates of solvency of a client.

Unlinered classification of model's logit and profit. In order to solve the task of classification of applicants for receiving the credit the function of division of probabilities (cumulative function of division (CFD)) changed in a necessary way is used. CFD belongs to the class of monotonous functions that means functions which increase and decrease in a monotonous way on a certain interval. Let's allow that in order to define the probability of receiving credit a normal division is chosen:

$$p_c = \Phi(b^T x) = \int_{-\infty}^{u} \varphi(z)\, d\, z$$

$\varphi(z)$ - density of a normal standard division; $u = b^T x$ - upper boundary of integration. Thus, in this way we can get a model called probity.

If in order to define the probability of receiving credit the function of logistics division is used, the unlinered model of logit can be built. In this case we have:

$$p_c = \Phi(b\ x) = \int_{-\infty}^{u} \varphi(z)\, d\, z = \frac{1}{1+\exp(-b^T x)} \quad (4)$$

$$p_c = \frac{\exp(b_1 x_1 + ... + b_m x_m)}{1+\exp(b_1 x_1 + ... + b_m x_m)}$$

In difference from the function of normal division logistics function can closed form that provides simplified calculation using this model in comparison with the model of probity. The parameters of both models are usually estimated using the method of the maximum plausibility (MMP) without calculation expenditures. The alternative method of estimation is Monte Carlo method for Markov chains (MCMC) which is based upon generating of pseudorandom sequences (GPS) and selection of casual values that correspond to certain demands. This method is widely used for the estimation of unlinered models because of alternative methods of generating GPS. For instance, it is used in the procedures of analysis of financial processes for the estimation of models of stochastic volatility [5]. The results of client's classification received from using both models in the majority of cases are accepted because of quality [1].

Discriminant analysis. Problem definition in this case also demands the division of clients in two groups: G_1 – the group of clients which correspond the requirements, and G_2 - the group which is not given the credit. The task is in classification potential clients on the bases of plural of their characteristics $x = [x_1, x_2, ..., x_m]$. The technique of discriminant analysis solves this task by calculation of discriminant function $\lambda^T x$, λ is a function of weight coefficients for component vector x. The meaning of weight coefficients is calculated by defining the maximum possible difference between both groups of clients.

It is thus allowed that vector x has normal division for both groups. Each group can be placed in correspondence to several meanings (parameters) (μ_1, Σ_1) i (μ_2, Σ_2) which represent group average and covariance accordingly. Also it is necessary to put in probability p_i - a probability of accessory of a possible credit owner to a group i, and a quantity c_{ij} which characterizes expenditures connected with wrong classification (when a credit applicant should be placed from group i to group j). In case when covariance matrixes of both groups are the same meaning $\Sigma_1 = \Sigma_2 = \Sigma$ then the rule of classification can be received from the condition of minimizing of the cost of expected wrong classification. Such interpretation leads to such a result: a credit applicant which is characterized by a plural of meaning $\mathbf{x}$ will be put to group G_1 if:

$$\lambda^T x \geq \alpha + \ln\left(\frac{c_{21}\, p_2}{c_{12}\, p_1}\right),$$

where

$$\lambda = \Sigma^{-1}(\mu_1 - \mu_2); \quad \alpha = \lambda^T(\mu_1 + \mu_2)/2. \quad (5)$$

In other cases, an applicant is referred to group G_2

The rule of classification in this case is very simple: the meaning received with the help of discriminant function $\lambda^T x$ is compared with threshold which is defined by the following:

$$"threshold" = \alpha + \ln\left(\frac{c_{21}\, p_2}{c_{12}\, p_1}\right)$$

In case of exceeding the threshold, an applicant can be referred to the first group and on the contrary. Expression (5) is called the model of linear discriminant analysis because vector x enters it linearly. In case if

$\Sigma_1 \neq \Sigma_2$ the rule of classification will have square form concerning x; that is why such model is referred to square discriminant analysis.

The algorithm of recursive classification (ARC). The algorithm of recursive classification is based on the rule of sorting which uses the consequences of binary plurality of explanation criteria. The usage of ARC is resulted in binary classification tree knots and rods of which create a structure that makes a correspondence of a certain meaning of a classification group (G_1 or G_2) to the data x_i of a potential client. Let's make a simple illustration of this approach.

It is necessary to divide N subjects into two groups G_1 and G_2 using two criteria: A and B. Carrying to a group means minimization of the expected cost of wrong classification. In other words, it is necessary to minimize risks of a necessity to change applicant's accessory to a group. A risk of identification the final point t of a classification tree to a group G_1 can be formalized the following way:

$$R_1(t) = c_{21}\,\pi_2\,p(2|t) \qquad (6)$$

π_i Is a possibility of a subject's accessory to a group i; c_{ij} - the cost of an applicant's accessory to the credit of the group j if he belongs to the group i; $p(2|t)$ - conditional possibility of a thing that a subject who belongs to the group G_2 will be identified a final point of a tree t. According to the analogy the risk of identifying the final point t of a classification tree to the group G_1 can be described the following way:

$$R_2(t) = c_{12}\,\pi_1\,p(1|t).$$

Thus if $R_1(t) < R_2(t)$ then the algorithm will identify the last point t of a classification tree to the group G_1, and in other cases – to the group G_2.

The algorithm of recursive classification divides basic data (sub sampling) into two parts on the top of a classification tree. The data sorting is executed on the basis of concept of mixing of a sampling with the help of a chosen criterion or linear combination of several criteria (characteristics). As it is said above $p(2|t)$ is a conditional possibility of a fact that a subject who belongs to the group G_2 is identified to the last point of a tree t. On the whole, it is possible to define a conditional possibility $p(i|t)$ concerning group i.

Formally the level of mixing of the meaning identified to the last point t can be written the following way:

$$I(t) = R_1(t)\;p(1|t) + R_2(t)\;p(2|t)$$

The function $I(t)$ can be interpreted as an expected risk which appears because of the wrong classification that means in the final point t a subject is referred to both groups but in reality it is referred to the group i with a possibility $p(i|t)$. The integrated level of mixing $I(T)$ for the classification tree T can be defined as aggregation estimation for all final points.

It is obvious that the level of mixing of all samplings in any point of a tree will be bigger than the degree of mixing of several samplings taken from the whole sampling. Thus, it is logical to formulate such a rule of classification in a site t which will give the possibility to decrease the degree of data mixing. Considering the facts ARC finds first the better rule in the given point for each of the characteristics and their combinations and on the following basis creates sub samplings. Such a procedure of binary classification continues till the level of data mixing can be decreased. On this point the process of classification comes to its end and we get the classification tree $T_{\max}$.

The last step of ARC is to choose necessary level of a tree's difficulty using the method of cross checking. Very often the decision trees got by this algorithm have a high level of difficulty that's why the risk of incorrect classification can be huge. Thus, it is necessary to make checking of a built tree at least using a part of data got in the process of building this model. Practical usage of this method means its good classification characteristics despite high difficulty of a model. It is explained by its non-parametrical nature.

Mathematical programming. This method of classification belongs to non-parametrical methods. It gives more possibilities for practical usage than parametrical statistics models which usage is limited by the meanings of parameters' estimation. Let's consider the task of dividing data into two groups: G_1 and G_2. The meanings of variable and criteria of classification for i – subject is in vector A_i. The task is to define such a vector $\mathbf{x}$ and limitations b which correspond the condition:

$$A_i x \le b, \text{ if } A_i x \in G_1;$$

$$A_i x \ge b, \text{ if } A_i x \in G_2.$$

Two groups are divided by a hyper square $Ax = b$. If to define through α_i the level of violation of this condition by a subject which is characterized by a data

vector A_i , to solve the task of classification it is necessary to find:

$$\min \sum_i c_i \alpha_i$$

where

$$A_i x \le b + \alpha_i, \text{ if } A_i x \in G_1; \quad A_i x \ge b - \alpha_i, \text{ if } A_i x \in G_2. \tag{7}$$

Bayesian networks (BN). Bayesian network is a model of a conditional type in a form of the directed acyclic count the tops of which are chosen variables of a process that is modeled. Each top is put in compliance with a table of conditional possibilities which is necessary for the calculation of future states of the top. The goal of building up such a model is to establish investigative connections between variables to get a possibility of forming conditional conclusion that means a conditional possibility of the events which we are interested in a concrete case.

In order to create a model in the form of BN it is necessary to solve tasks of structural and parametrical education that means this is a classic task of mathematical modeling of processes of any nature. In the process of BN building a priory structure of a network can be set empirically that means to get it with the help of expert estimation or other information concerning the investigating process. If a structure is unknown it is estimated with available information. In case if statistic information is available it is convenient to use heuristic algorithm of BN building that will correct a priory structure of a network or will give a possibility to build such a structure of a network that would be easy to modify later using the expert knowledge [6,7]. The result of BN usage is the calculation of a possibility of unreturning of a credit on those conditions which correspond the meaning of other model's variables. In table 1 there are the results of usage of three methods before the analysis of a borrower's solvency (bank data).

Table 1

Comparative table of received results of the models

Name of method	GINI Index	AUC meaning	Accuracy of model	Quality of model
Binary logistic regression	0,669	0,828	**0,776**	high
Decision trees	0,612	0,766	0,754	acceptable
BN	**0,687**	0,845	0,757	high

The note: AUC (Area Under Curve) is calculated for example with the help of trapeze method:

$$AUC = \int f(x)dx = \sum_i \left[\frac{X_{i+1} + X_i}{2} \right] \cdot (Y_{i+1} - Y_i)$$

The general accuracy of a model is defined as the relation of correctly forecasted cases to their total quality.

The received results mean that the best models of estimation of a borrower's solvency are those that are built by the method of logistic regression and BN. The best accuracy is also received with the help of logistic regression. These results confirm once more the expediency of using logistic regression and decision trees in the process of estimation of a borrower's solvency.

The method on the basis of inner credit rating. The method on the basis of inner credit rating (ICR or Internal Rating Based Approach – IRB Approach) is a leading method of estimation of credit risks [3, 4]. It gives a possibility to create flexible mechanisms of measuring of expected and unexpected expenditures. With the help of this method individual and group credit risks can be estimated. The main indicators that characterize the volume of potential losses using ICR are: 1) – the probability of default (PD) which gets the meaning from 0 to 1; 2) – credit exposure (CE) – the sum of credit debt; 3) – loss given default (LGD); it gets the meaning from 0 (the credit is completely covered by deposit) to 1 (the credit is completely uncovered by deposit); 4) – maturity M.

All the investigations in the process of evaluation of credit risks are conducted in the direction of creating a mechanism of calculation a possibility of borrower's default. In order to estimate the possibility of default it is necessary to build up the mechanism of this estimation. Problem definition of estimation of default possibility concerning individual credit risk can be formulated in the following way: on the basis of borrower's parameters and the meaning of the credit x_i^j it is necessary to create the procedure of estimation the probability of default PD_i :

$$PD_i = F(w^j, x_i^j) \tag{8}$$

w^j - Weight coefficient for x^j parameters. In order to solve the task two approaches can be used: 1 – scoring approach to build up mathematical model on the basis of default statistics for the former periods; 2 – expert method. The example of the usage of this approach to calculate the volume of expected losses by the portfolio that consists of 10 borrowers is illustrated in table 2.

Table 2

The example of calculation of portfolio expected losses

Borro-wer	CE (thou-sand hryvna)	Deposit (thou-sand hryvna)	LGD	Borro-wer's PD	Borro-wer's ICR
1	220	200	0,05	0,0002	AAA
2	60	40	0,45	0,015	A
3	250	180	0,2	0,033	BB
4	120	20	0,95	0,045	B
5	50	0	0,90	0,022	A
6	250	100	0,45	0,018	BBB
7	140	80	0,85	0,0045	AA
8	200	20	0,88	0,017	AA
9	20	3,0	0,02	0,024	A
10	50	5,0	0,99	0,015	BBB

The note: AAA = 0,0001; AA = 0,005; A = 0,01; BBB = 0,02; BB = 0,03; B = 0,05; CCC = 0,1; CC = 0,25; C = 0,5; D = 1.

The volume of expected credit losses for this portfolio makes 18,7 thousand hryvna. Thus, this approach gives a possibility to estimate the volumes of possible losses for the groups of borrowers simultaneously.

IV. CONCLUSIONS

Modern approaches to the solution of crediting tasks on the condition of minimizing risks of possible losses demand implementation of new effective principles of managing risks and computer systems of decision making support. To build up such systems it is necessary to develop and use plurality of alternative methods of data analysis, alternative models and certain criteria of analysis of model's quality and final result – a possibility of unreturning a credit.

In the Paper, there has been executed the analysis of some modern approaches to create classification mathematical model's total usage of which will give a possibility to make correct reasonable decisions concerning giving out credits to the clients of financial institutions. The best results of clients' classification using actual statistic data is received by binary unlinered models and Bayer's network. This means that models of such type have better indicators of statistic parameters of quality. Also, it is perspective to further develop the method on the basis of inner credit ratings that provides complete information concerning the situation with credits. For example, it is possible to get the estimation of possible losses.

In further research, it is necessary to improve chosen types of classification models in order to increase the quality of clients' classification into two groups. Also, it is reasonable to use simultaneously «ideologically» different types of models – regression, conditional, neuro-network and neuro-illegible.

REFERENCES

[1] BIDIUK P.I., MATROS YE. O..: Modeli otsiniuvannia ryzykiv kredytuvannia fyzychnykh osib. Kibernetika ta obchysliuvalna tekhnika.- 2007.- N. 153.- S. 87 - 95.

[2] Pod. redakciej LOBANOVA O.O., CHUGUNOVA A.B.: Enciklopediya finansovogo risk-menedzhmenta. M.: Alpina Pablicher, 2003.- 845 s.

[3] KISS F. Credit scoring processes from a knowledge Management perspective. Hungary Periodica Polytechnica. - 2003. - vol. 11, ? 1. - P. 95-110.

[4] JORION P.: Financial risk management handbook. New Jersey: John Wiley & Sons, Inc., 2003. – 422 p.

[5] BIDIUK P.I., KONOVALIUK M.M.: Otsiniuvannia modelei stokhastychnoi volantylnosti ta UARUH na Java. Naukovi pratsi: Kompiuterni tekhnolohii.- Mykolaiev: ChDU im. P. Mohyly, 2012.- Vyp. 179, t.191.- S. 14 - 20.

[6] BIDYUK P.I., TERENT'EV A.S., GASANOV.: Postroenie i metody` obucheniya bajesovskix setej. Kibernetika i sistemny`j analiz 2005.- N 4.- S. 134 - 147.

[7] COWELL R.G., DAWID A.PH., LAURITZEN S.L., SPIEGELHALTER D.J.: Probabilistic networks and expert systems. New York: Springer, 1999. - 323 p.