Credit scoring model for microfinance organizations

Svitlana O. Yaroshchuk, Nonna N. Shapovalova[0000-0001-9146-1205], Andrii M. Striuk[0000-0001-9240-1976], Olena H. Rybalchenko[0000-0001-8691-5401], Iryna O. Dotsenko[0000-0001-7912-2497] and Svitlana V. Bilashenko[0000-0002-4331-7425]

Kryvyi Rih National University, 11, Vitalii Matusevych Str., Kryvyi Rih, 50027, Ukraine
yaroschucksvetlana@gmail.com, shapovalovann09@gmail.com, andrey.n.stryuk@gmail.com, ellinaryb@gmail.com, irado441@gmail.com, SvitlanaViktorivnaBilashenko@gmail.com

Abstract. The purpose of the work is the development and application of models for scoring assessment of microfinance institution borrowers. This model allows to increase the efficiency of work in the field of credit. The object of research is lending. The subject of the study is a direct scoring model for improving the quality of lending using machine learning methods. The objective of the study: to determine the criteria for choosing a solvent borrower, to develop a model for an early assessment, to create software based on neural networks to determine the probability of a loan default risk. Used research methods such as analysis of the literature on banking scoring; artificial intelligence methods for scoring; modeling of scoring estimation algorithm using neural networks, empirical method for determining the optimal parameters of the training model; method of object-oriented design and programming. The result of the work is a neural network scoring model with high accuracy of calculations, an implemented system of automatic customer lending.

Keywords: neural network, machine learning, lending, scoring.

1 Introduction

Increasing the profitability of credit operations is directly related to the quality of credit risk assessment [1]. In recent years, there has been a rapid increase in retail lending. Competition is increasing, the range of services is expanding, the process of obtaining a loan is being simplified, and the decision-making time is significantly reduced. The quality and speed with which a credit request is generated, as well as the reliability and simplicity of this process are crucial factors in a complex competitive process.

An important component of the bank's stable development under conditions of volatile financial position is the compliance of the risk management system with modern standards of quality of management, as well as the degree of protection against unpredictable external influences. Thus, banking organizations need to introduce scoring systems that allow to resolve qualitatively emerging issues.

Scoring is a way of quickly evaluating a potential customer of a bank or microfinance organization. The assessment is performed by analyzing the borrower's questionnaire

and calculating each customer's score according to the rules set out in the specific financial structure. The scoring system is a special program with a built-in algorithm for deciding on certain parameters. The reliability and quality of the response depends on the quality of the algorithm. Therefore, it is important not only to create a scoring model, but also to minimize the error of results.

2 Research apparatus

The aim of the study is the theoretical justification and development of a scoring model for microfinance borrowers.

Objectives of the study:

1. To analyze the needs of the lending industry in applying scoring assessment, types of scoring and optimal solutions for this sector.
2. To consider methods of constructing a model for scoring assessment, choose the most optimal one.
3. To choose a model architecture, create software for practical demonstration of scoring.
4. To process the input data, evaluate the initial result and achieve maximum accuracy of the system.

The object of research is creating software for scoring of borrowers.

The subject of research is the development of a scoring assessment model for microfinance organizations.

Research methods: analysis of the literature on banking scoring; artificial intelligence methods for scoring; modeling of scoring estimation algorithm using neural networks, empirical method for determining the optimal parameters of the training model; method of object-oriented design and programming.

The practical significance of the obtained results is a software for microfinance organizations that helps to assess the risk of issuing a loan to a specific borrower, thereby improving the efficiency of these institutions.

3 Theoretical foundations of banking scoring

Credit is an important category of a market economy that reflects the real ties and relationships of economic life in society. The loan originated from the practical needs of production development, its adaptation to the conditions of permanent capital shortage – monetary and material resources.

Credit relations operate in the system of economic relations. They are based on the movement of a special kind of capital – loan capital.

In today's context, the approach to credit organization has changed fundamentally: there has been a shift from object to direct lending to entities. This means that the emphasis in the lending mechanism has shifted from the selection of the entity to the entity's valuation. Commercial and partnership relations between the parties to the

agreement exclude the creditor's dictate in determining the object of credit. The risky operations that give the highest income to the bank need to study not only the effectiveness of the activities (projects) under which the funds are allocated, but also the creditworthiness of the client.

A borrower's creditworthiness is his ability to fully and timely settle financial obligations.

The borrower's creditworthiness, unlike its solvency, does not record any insolvency for the current period or for any date, but predicts its solvency in the near term.

One way to organize credit relations is to qualitatively assess the creditworthiness of the borrower. Commercial banks are in dire need of information about the creditworthiness of farms. Their profitability and liquidity depend largely on the financial position of the customers, since the reduction of the risk when performing loan operations can be achieved only based on studying the creditworthiness of clients.

One of the most effective tools for such assessment is the scoring system. Credit scoring enables to make a quick and qualitative decision on a loan application. In addition, its reliability and simplicity are crucial factors in the complex competition [10].

In general, credit scoring can be defined as an assessment of the level of credit risk that results from the processing of various credit history data, which directly or indirectly affects the level of payment discipline [1].

Applying credit scoring, that is, a systematic approach to dealing with credit applications as a whole, allows the bank to:

- Increase the loan portfolio by reducing the number of unjustified refusals of loan applications;
- Improve the accuracy of the borrower's valuation;
- Reduce the level of defaults;
- Speed up the borrower's valuation process;
- Create centralized accumulation of borrower data;
- Reduce provisions for possible losses on credit liabilities;
- Quickly and qualitatively evaluate the dynamics of changes in the credit account of the individual borrower and the credit portfolio as a whole.

All of these have many advantages over a conventional customer rating system and establish the bank's performance.

4 Definition of creditworthiness of the client

Determining the borrower's creditworthiness is an important step in approving a loan application. The main task of the lender is to assess all the risks associated with the possibility of non-repayment of the funds provided.

Credit companies consider the following nuances when evaluating creditworthiness [3; 10; 15]:

- The financial position of the potential borrower;

- Debt load – the ratio of existing liabilities and the requested loan to the applicant's principal income;
- Credit reputation of the client;
- The value of the property owned by the borrower;
- The applicant's social status, personality, career advancement and other factors.

Banking organizations analyze the ability of an individual to pay on a loan. In this case, not only the borrower's monthly income and expenses, but also other factors are taken into account. For example, the risk of job loss and other insured events [15].

Microfinance companies use cheap and fast valuation methods. Therefore, it will take a minute to conduct a scoring test. This is very handy for small loans. However, when it comes to large bank loans, credit professionals can apply all of these methods in combination. This approach will make a specific prediction.

5 Data and methods of scoring in microfinance

5.1 Structure of the main modules

Designing a system that solves the problem of credit scoring can be divided into two main modules: data processing, which includes bringing data to a format that is favorable for computer computing, and directly the module of calculations of the loan decision, containing the interpretation of the algorithm selected for solving the problem teaching. The output of the first module is the input for the second. Therefore, the quality of the result depends on the degree of processing of the primary data.

Data processing includes:

- Data preparation: delete duplicate records, non-informative data columns, records with many null values. Assess the significance of each trait included in the training sample by conducting correlation and regression analyzes.
- Data conversion: categorical features are reduced to a vector form and numeric values should be reduced to a single standard, such as the interval [0, 1] or [-1, 1].

The calculation module consists of:

- Choosing the architecture, parameters of the algorithm.
- Model training on the selected algorithm.
- Testing the model on a deferred sample, determining the calculation error with further correction [9].

5.2 Methods of scoring

Credit scoring is a typical machine learning task. It refers to the type of supervised learning (training with the teacher) [4], namely to the problems of classification, because the solution of the task is reduced to the identification of risks of granting credit for two types (classes): "good" and "bad". The good ones will be those customers who are likely to repay the loan, the bad ones – those who will have a delay of more than 3

months. Sometimes, the “bad” risks include those customers who repay loans early or within a specified period, and the bank does not have time to profit from such clients.

The credit scoring problem can be solved by different machine learning classification methods [7]. These include [2]:

- Statistical methods based on discriminant analysis (linear regression, logistic regression);
- Different linear programming options;
- Classification tree or recursion-partition algorithm;
- Neural networks;
- Genetic algorithm;
- Method of nearest neighbors.

Traditional and most common are regression methods, primarily linear multivariate regression [2]. The disadvantage of the model is that on the left side of the equation is a probability that takes values from 0 to 1, and variables on the right can take any values from $-\infty$ to $+\infty$. In addition, this model is unstable to emissions; any sudden value that gets out of the picture can lead to the wrong answer.

Linear programming also leads to a linear scoring model. It is impossible to carry out a completely accurate classification of “bad” and “good” clients, but it is desirable to minimize the error. The task is to find the weights for which the error will be minimal [17].

Classification trees are a method that allows the observation or object to be assigned to a particular class of categorical dependent variable according to the values of one or more predictor variables. Classification trees are tailored to the graphical representation, so they have a more convenient look for human understanding. The disadvantages are instability, small changes in the data can significantly change the built decision tree, the problem of finding the optimal depth of the tree, the complexity of data gaps.

The genetic algorithm is based on an analogy with the biological process of natural selection. In the field of lending, it looks like this: there is a set of classification models that can be “mutated”, “crossed”, and as a result, the “strongest” model is selected, which gives the most accurate classification.

When using the nearest-neighbor method, a unit of measure is selected to determine the distance between clients. All clients in the sample are given a specific spatial position. Each new client is classified based on which clients – good or bad – are more around him [2].

Neural Networks – a common solution to classification problems. Artificial neural network – a mathematical model, which is built on the principle of organization and operation of biological neural networks, is a system of connected and interacting simple processes [8].

In scoring, the use of neural networks has the least use compared to other methods. Nevertheless, the neural network has significant advantages. These advantages include the possibility of automatic learning of the model, the versatility of working with different scales of measurement of dependent and independent variables, the ability to approximate any continuous function of dependence. The mathematical model of

neural network scoring makes it possible, based on a set of known characteristics of a research object, to predict a specific characteristic that is unknown to the researcher [6; 18].

The neural network model meets all the above-mentioned needs of the domain, such as speed and precision of calculations, scalability of data, possibility of introduction of new characteristics without significant deterioration of quality of estimation, easy modernization of the system on demand. Therefore, based on the advantages of using neural networks, this method was chosen to solve the credit-scoring problem.

Building a neural network has its own characteristics and steps that you must go through to get a truly high-quality model.

Building a neural network starts with data preparation. At this point, it is necessary to delete duplicate records, non-informative data columns, records that have many null values. At the same time, there should be a sufficient number of examples for neural network training. There is an imperial rule that establishes the recommended ratio between the number of training examples that have input and the correct answers and the number of connections in the neural network: $X<10$.

For the facts included in the training sample, it is advisable to estimate its significance in advance by conducting correlation and regression analyzes and to consider the ranges of their possible changes. All this is an important component in the methodology of building a scoring model [3].

The second stage is the conversion of the initial data, taking into account the nature and type of problem solved by the neural network, the means of presentation of information are selected. For example, categorical features should be reduced to a vector form and numerical values should be reduced to a single standard, such as the interval [0, 1] or [–1, 1].

The third stage is the choice of network architecture. It is necessary to define such parameters as the number of network layers, the number of neurons of each layer, and the activation functions to be used [18].

The number and type of independent variables in the model determine the number of neurons in the input layer. For categorical variables, it is advisable to use one input neuron for each category, with only one neuron in the group being activated for each observation.

The architecture of the source layer of the neural network is also dictated by the structure of the problem. One output neuron is created for each dependent quantitative variable.

No techniques have been developed to determine the number of hidden layers and the number of neurons in them. In practice, these parameters are experimentally determined by analyzing the quality of approximation provided by networks of different sizes. Thus, a network with an input layer of 58 inputs, one hidden layer with 30 neurons and 2 outputs was selected.

At the input of a neuron, we have a vector of parameters. These are the results of the collection of billing information about a potential customer, presented in numerical form $X^i = \{x_1^i, x_2^i, \dots, x_n^i\}$.

In this case, each client is responsible for the class Y^i. In total, there are two classes of the set Y with the following values: 1 – to give credit, 0 – not to give. The neural

network, in fact, must find the optimal separating hypersurface in the vector space, the dimension of which will correspond to the number of features. Learning the neural network in this case is to find such values (coefficients) of the weight matrix, in which the neuron responsible for the class will give values close to one in cases where credit is approved, and values close to zero, if not.

As we can see from formula 1, the result of the neuron $h_w(X)$ is a function of activation f from the sum of the product of the input parameters x_k to the coefficients required $w_k \cdot x_k$ in the learning process:

$$h_w(X) = f\left(\sum_{k=1}^{|w|} w_k \cdot x_k\right) h_w(X) = f\left(\sum_{k=1}^{|w|} w_k x_k\right) \quad (1)$$

It is desirable to interpret the value derived from a neuron in the range [0, 1] as the probability of belonging to a class. Therefore, such a monotonous smooth function is required, which will display elements of the set of real numbers in the range from zero to one. The activation function of sigmoid (2) is the best way to do this. The function graph is shown in Figure 1.

$$\partial(w \cdot x) = \frac{1}{1+e^{-x}} \quad (2)$$

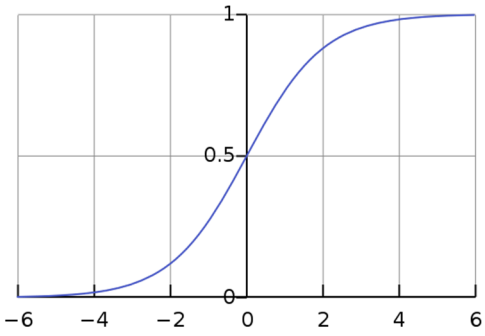

Fig. 1. Plot of sigmoid function

The fourth stage is learning the network. In the selected data set, each sample object is assigned a class to which it belongs. Therefore, we are tasked with the type of supervised learning [16]. Therefore, in the learning process, the network must review the sample many times, each complete passage of the sample is called the learning age. To train the model, you need to split the data into two parts – the actual training and the test.

We apply the standard separation in the ratio of 80% and 20% respectively [11].

Neural network training should be understood as the weighting for each of the traits based on the results obtained from past data views. The backpropagation method is selected for weight correction. See [5] for more on this method.

It has been determined experimentally that 64 epochs are required to select the optimal weights for a given neural network. 221,712 training sample records are processed for each epoch. The results are Table 2.

Table 1. The results of calculating epochs

Epoch number	Calculation time, s	Error	Accuracy
1	22	0.4871	0.6860
2	21	0.3066	0.7271
3	22	0.2472	0.7748
4	21	0.2128	0.7951
5	23	0.1970	0.8155
…	…	…	…
61	22	0.0304	0.9518
62	22	0.0275	0.9529
63	22	0.0255	0.9537
64	22	0.0251	0.9540

The fifth step is to test the received neural network model on a delayed sample. The test sample includes only those records that did not participate in the training network. We have 55,428 records. The accuracy of the calculations on the test sample is 95.4%. In this case, the error in the decision to issue a loan – 0.00469, the error in the loan prohibition decision is 0.00406.

6 Software architecture and operation

6.1 Architecture

The client server architecture of the program was chosen to develop the software for rapid assessment of customer solvency.

Client-server is a software architecture model of two parts, client systems and server systems, both communicating over a computer network or on the same computer. A client-server application is a distributed system made up of both client and server software. The client-server application provides a better way to share the workload. The client process always initiates a connection to the server, while the server process always waits for requests from any client.

The client-server relationship describes the relationship between the client and how it makes a service request to the server, and how the server can accept these requests, process them, and return the requested information to the client [12].

6.2 Description of software operation

To build a functional diagram, the IDEF0 methodology was chosen, which is considered the classic method of the process approach to design. In IDEF0, the system that is being modeled is represented as a set of interrelated works (functions, activities). To develop a functional diagram, 4 main functions of the program were presented, which are presented on blocks A1–A4 (Fig. 2). Each of these functions has input and output data, control information and mechanisms through which the function can be executed.

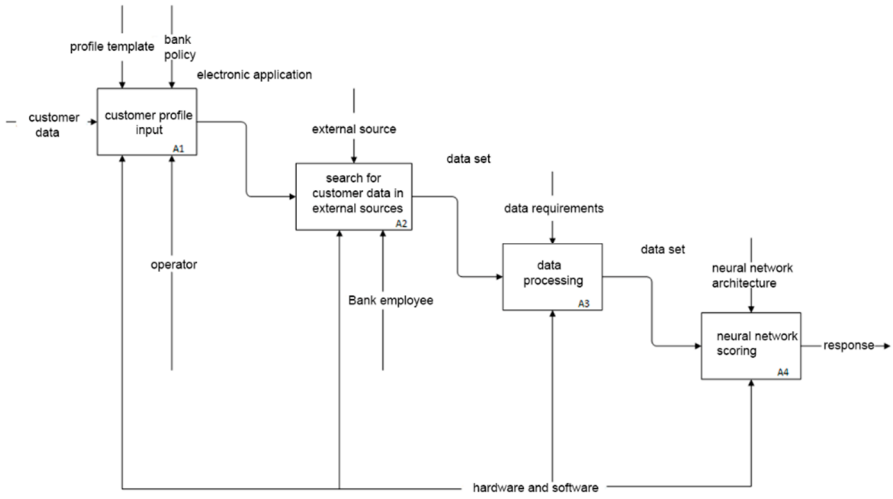

Fig. 2. Functional program scheme

Block A1. Block maintaining the customer profile. At the input, it has data coming from a bank client. This data is written to the database in a format structured in accordance with the bank policy and the accepted questionnaire template. The bank operator maintains the questionnaire. It can both enter data, view it, perform a search, delete it.

Block A2. After the data are entered into the database and the electronic application is generated, the customer data is searched in external sources. As a rule, these are Internet portals or credit bureaus. This approach helps to obtain more detailed information about the client and his solvency.

Block A3. Data processing. It provides functionality for preparing data for further calculations, all non-informative data is deleted, data is brought to a single range. Data processing is based on the analysis of data requirements.

Block A4. The processed network dataset enters the neural network model. The structure and quality of functioning of the neural network are determined by its architecture. The model gives the result of a loan to the borrower.

All blocks as the executing mechanism have software and hardware.

The system is a module for the banking system, so this imposes certain requirements on the program interface.

Firstly, an official and minimalistic style of design should be maintained. Light background colors of the components and dark text color on them.

Secondly, all elements should be located at an optimal distance from each other, to prevent errors that in the banking sector can cost both time and money.

Thirdly, the size of the inscriptions should be sufficient for the readability of the text. The text on the components must match the actual functionality that the component executes.

An important part of the interface is the presence of messages about the actions performed in the program, as well as the status of request processing.

The developed software consists of several tabs: “Clients”, “Data Analysis” and “Settings”.

The Customers tab has buttons for managing customer data. It is possible to add new clients, remove them from the database, search the client for the database, and directly score on the selected client.

The developed software consists of several tabs: “Clients”, “Data Analysis” and “Settings”.

The Customers tab (Fig. 3) has buttons for managing customer data. It is possible to add new clients, remove them from the database, search the client for the database, and directly score on the selected client.

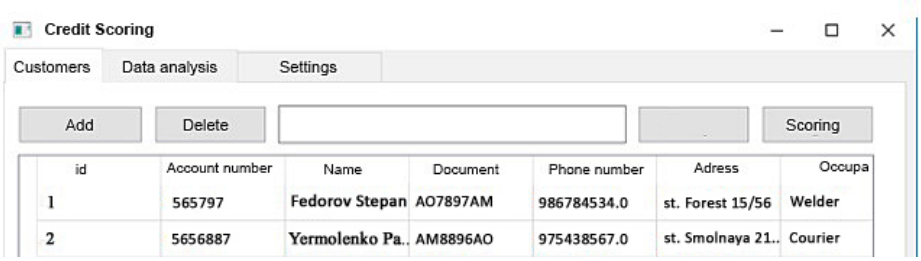

Fig. 3. Customers tab

To view the analytics data to monitor the status of credit disbursements and other parameters, go to the Data Analysis tab (Fig. 4). The graph will be automatically generated according to the selected criterion from the drop-down menu.

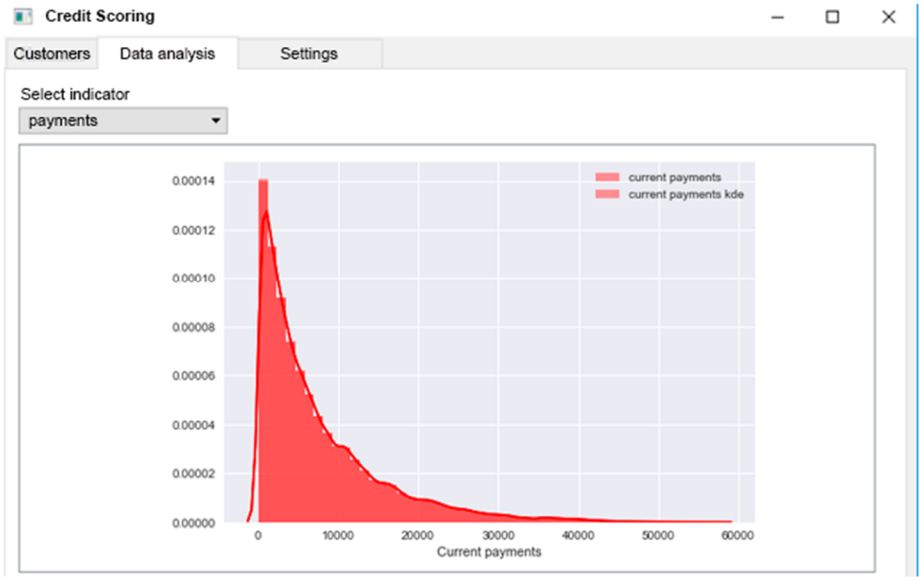

Fig. 4. Data Analysis Tab

The user can also change the neural network settings by going to the Settings tab. The banking system is dynamic and sometimes it is necessary to re-evaluate the weight of the features. In the list of features, you should select the ones that will be included in the new model using the multiple choice in the box. You can select the number of hidden layers and the number of neurons in them.

7 Results

As a result, a system of automatic crediting of clients in the sphere of microfinance was created. For this system, the best artificial neural network architecture with all the relevant settings was selected.

Using this decision support system on the artificial neural network platform to make serious decisions such as credit decisions significantly simplify lending operations, reduce the risk of default.

Testing the system gave fairly accurate results, which indicates a high degree of trust in the software.

The result of the program (Fig. 5):

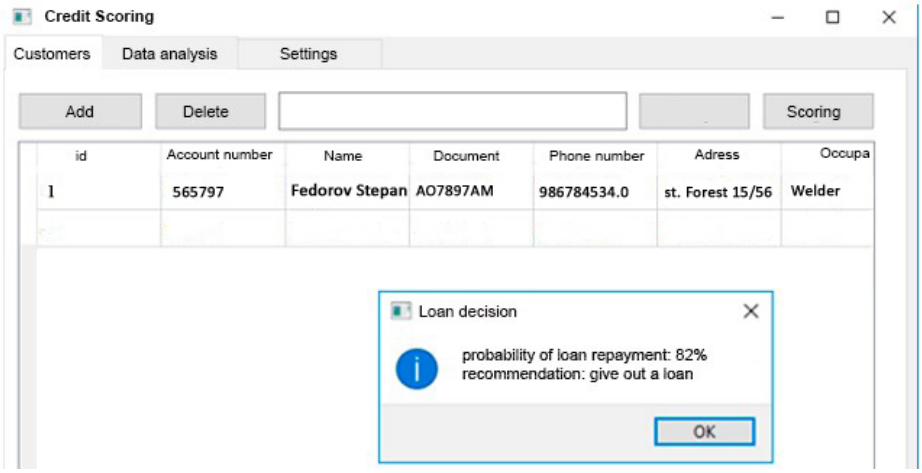

Fig. 5. The result of the program

8 Conclusion

The study showed how artificial neural networks could be used to build an automatic customer lending system.

In accordance with this goal, the following results were obtained: the current state of the problem and task of crediting clients are analyzed; the modern decision support systems are analyzed and the choice of artificial neural network systems as the basic technological platform is grounded; the data model necessary for the proper functioning

of the system is built; selected the best neural network architecture for a specific task; automatic system of crediting of clients is implemented.

Building a neural-boundary model for the credit-scoring problem has shown that this method is one of the best for finding the most accurate solution for issuing a loan, as evidenced by a very low calculation error.

The introduction of such a system in a microfinance institution helps to improve the performance of the institution, automate the processes associated with credit loans, reduce the risk of errors and fraud.

References

1. Aleshin, V.A, Rudayeva, O.O.: Kreditnyj skoring kak instrument povysheniya kachestva bankovskogo risk-menedzhmenta v sovremennyh usloviyah (Credit scoring as an instrument for improving the quality of banking risk management in current conditions). Terra economicus. **10**(2), 27–30 (2012)
2. Allison, P.D. (ed.): Logistic regression using the SAS system: theory and application. SAS Institute, Stanford (2012)
3. Anderson, R.: The credit scoring toolkit: theory and practice for retail credit risk management and decision automation. Oxford University Press, New York (2007)
4. Coelho, L.P., Richert, W.: Building Machine Learning Systems with Python. Packt Publishing, Birmingham (2013)
5. Flach, P.: Machine Learning: The Art and Science of Algorithms that Make Sense of Data. Cambridge University Press, Cambridge (2012)
6. Haykin, S.: Neural Networks and Learning Machines, 3rd edn. Pearson, New Jersey (2008)
7. Kiv, A., Semerikov, S., Soloviev, V., Kibalnyk, L., Danylchuk, H., Matviychuk, A.: Experimental Economics and Machine Learning for Prediction of Emergent Economy Dynamics. In: Kiv, A., Semerikov, S., Soloviev, V., Kibalnyk, L., Danylchuk, H., Matviychuk, A. (eds.) Experimental Economics and Machine Learning for Prediction of Emergent Economy Dynamics, Proceedings of the Selected Papers of the 8th International Conference on Monitoring, Modeling & Management of Emergent Economy (M3E2 2019), Odessa, Ukraine, May 22-24, 2019. CEUR Workshop Proceedings **2422**, 1–4. http://ceur-ws.org/Vol-2422/paper00.pdf (2019). Accessed 17 Aug 2019
8. Lewis, E.M. An introduction to credit scoring. Athena Press, London (1992)
9. Luo, F.L., Unbehauen, R.: Applied Neural Networks for Signal Processing. Cambridge University Press, Cambridge (1997)
10. Mays, E. (ed.): Handbook of credit scoring. Global Professional Publishing, Chicago (2001)
11. Rojas, R.: Neural Networks: A Systematic Introduction. Spring-Verlag, Berlin (1996)
12. Saternos, C.: Client-Server Web Apps with JavaScript and Java. O'Reilly Media, Sebastopol (2014)
13. Semerikov, S.O., Teplytskyi, I.O., Yechkalo, Yu.V., Markova, O.M., Soloviev, V.N., Kiv, A.E.: Computer Simulation of Neural Networks Using Spreadsheets: Dr. Anderson, Welcome Back. In: Ermolayev, V., Mallet, F., Yakovyna, V., Kharchenko, V., Kobets, V., Korniłowicz, A., Kravtsov, H., Nikitchenko, M., Semerikov, S., Spivakovsky, A. (eds.) Proceedings of the 15th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer (ICTERI, 2019), Kherson, Ukraine, June 12-15 2019, vol. II: Workshops. CEUR Workshop

Proceedings **2393**, 833–848. http://ceur-ws.org/Vol-2393/paper_348.pdf (2019). Accessed 30 Jun 2019

14. Semerikov, S.O., Teplytskyi, I.O., Yechkalo, Yu.V., Kiv, A.E.: Computer Simulation of Neural Networks Using Spreadsheets: The Dawn of the Age of Camelot. In: Kiv, A.E., Soloviev, V.N. (eds.) Proceedings of the 1st International Workshop on Augmented Reality in Education (AREdu 2018), Kryvyi Rih, Ukraine, October 2, 2018. CEUR Workshop Proceedings **2257**, 122–147. http://ceur-ws.org/Vol-2257/paper14.pdf (2018). Accessed 30 Nov 2018
15. Siddiqi, N.: Credit risk scorecard: developing and implementing credit scoring. John Wiley and Sons, New Jersey (2006)
16. Sorokin, A.S.: K voprosu validacii modeli logisticheskoj regressii v kreditnom skoringe (On the validation of the logistic regression model in credit scoring). Naukovedenie 2. http://naukovedenie.ru/PDF/173EVN214.pdf (2014). Accessed 10 Nov 2019
17. Sorokin, A.S.: Postroenie skoringovyh kart s ispolzovaniem modeli logisticheskoj regressii (Construction of scoring maps using a logistic regression model). Naukovedenie 2. http://naukovedenie.ru/PDF/180EVN214.pdf (2014). Accessed 10 Nov 2019
18. Sorokin, S.V., Sorokin, A.S.: Ispolzovanie nejrosetevyh modelej v povedencheskom skoringe (Use of neural network models in behavioral scoring). Prikladnaja informatika **10**(2(56)), 92–109 (2015)