МАТЕМАТИЧНІ МЕТОДИ, МОДЕЛІ, ПРОБЛЕМИ І ТЕХНОЛОГІЇ ДОСЛІДЖЕННЯ СКЛАДНИХ СИСТЕМ

UDC 004.5
DOI: 10.20535/SRIT.2308-8893.2025.3.08

MULTI-CRITERIA MATHEMATICAL MODEL OF CREDIT SCORING IN DATA SCIENCE PROBLEMS

O.O. PYSARCHUK, M.D. VASYLIEVA, D.R. BARAN, I.O. PYSARCHUK

Abstract. A multi-criteria optimization mathematical model of credit scoring is proposed. The model is derived using a nonlinear trade-off scheme to solve multi-criteria optimization problems, allowing for the construction of a Pareto-optimal solution. The proposed approach forms an integrated assessment of a borrower's creditworthiness based on a structured set of indicators that reflect the financial, credit, and social profile of clients. The model is designed for use in intelligent CRM and ERP systems operating on Big Data and does not rely on labeled training samples, making it applicable to unsupervised learning tasks. It can also serve as a foundational layer for further deep-learning analysis. Methodological steps for implementing the model, from indicator normalization to final decision-making, are described. A technological implementation demonstrates the model's effectiveness in automated loan decisions and fraud detection.

Keywords: Data Science, Big Data, SCORIG machine learning, decision making, multi-criteria mathematical models, intelligent CRM, ERP systems.

INTRODUCTION

The development of the modern IT industry determines the methodologies and technologies of electronic banking. This also affects the automation of intelligent decision-making processes. One of these directions is making decisions about granting loans to consumers of lending services (clients) provided by banking institutions. This process relates to the field of credit scoring (SCORIG). It is based on the analysis of a set of indicators of the client's creditworthiness and establishing an individual integrated assessment (SCORE) in order to make an informed decision on granting a loan. The practice of scoring analysis is not limited to a binary yes/no assessment of lending. Scoring analysis should ensure the formation of an adequate risk assessment that determines a specific credit program adapted in the loan life cycle to the properties of a specific client. The process of scoring analysis should ensure high economic performance indicators of the banking institution. For lending programs, this means maximizing the number of loans issued, but through programs that are adequate to the risks of non-repayment of loan funds by the client. Currently, credit scoring is implemented by automated software tools that have the properties of intelligence and are organized in the format of distributed CRM or ERP systems.

Features of credit scoring in modern CRM, ERP are, in fact, the implementation of Data Science technologies on Big Data arrays. This imposes rather strict requirements on the computational complexity of credit scoring models.

Undoubtedly, the quality of automated credit scoring decision-making is determined by the mathematical models underlying the automated software tools. Therefore, the pragmatic effort of banking institutions to increase the economic efficiency makes relevant the task of developing effective mathematical models of credit scoring in Data Science tasks on Big Data arrays.

ANALYSIS OF RESEARCH AND PUBLICATIONS

The specificity of the credit scoring task consists in considering it within the class of classification methods and models. Classical methods of machine learning are most often used [1–5]: discriminant analysis; logistic regression; decision trees; method of support vectors; naive Bayes classifier; neural networks and others. In general, the assessment task refers to the theory of system analysis [6]. Performance evaluation is carried out in the following sequence: determining factors, indicators and criteria; forming the decision-making model; interpreting the obtained result. Single-criteria and multi-criteria models of efficiency evaluation are distinguished [7, 8], with the latter being comparatively more adequate. The main drawback of traditional credit scoring approaches is the consideration of the assessment task at the level of classification and forming an integrated assessment on a discrete field of static numerical representations of many factors of a particular borrower.

Currently, digital twin technologies are rapidly developing, which involve implementing any business processes into digital virtual reality in order to automate and optimize them. Digital twin is effective not so much in the aspect of automation as in the maximum approximation of the “twin” to the real physical process. This means that the twin must have a high level of adequacy but with abstraction that allows productive processing of, for example, Big Data arrays.

The main idea of the approach proposed in the article is to approximate the problem of credit scoring to its real physical essence at the formalization level. This means that a real client should come as close as possible to the image of an unattainable ideal based on a set of indicators. In this formulation, the task of scoring analysis reflects the task of multi-criteria evaluation. This will be applied to the synthesis of a mathematical model of scoring analysis, as a multi-criteria optimization mathematical model of evaluation.

Multi-criteria formalization of the decision-making task a priori has a higher level of adequacy since it mathematically allows describing a specific practical task of natural language expressed by the scheme “how best...”. Moreover, the criterion allows describing the entire set of possible indicator values, even though represented by a set of discrete limited realizations.

Moreover, the criterion allows describing the entire set of possible values of indicators, although represented by a set of discrete limited implementations. Thus, the increased adequacy of the evaluation task at the formalization stage ensures that the “twin” closely approximates the real physical process. Therefore, we should potentially expect an increase in the efficiency of the final result of the scoring analysis.

Formulation of the problem. The aim of this article is to synthesize a multicriteria optimization mathematical model for credit scoring.

PRESENTATION OF THE MAIN MATERIAL

The achievement of the stated goal is implemented at three levels: model, methodological, and technological.

I. *Multicriteria Optimization Mathematical Model of Credit Scoring*. The synthesis of the mathematical model is implemented in stages: defining factors, indicators, and criteria; forming the decision-making model; interpreting the obtained result.

Defining factors, indicators, and criteria are the initial data for evaluation and represent the scoring card. The scoring card structure is known but may vary in the number and values of indicators according to the specific conditions of a particular banking institution.

The classical structure of the scoring card includes the following groups of indicators[1–5]: information from the banking institution—credit product; information about the borrower/client—credit history; financial; social (see Table 1).

Table 1. Scoring card — general structure

Credit product (bank): ▪ Amount; ▪ Term; ▪ The purpose of the loan; ▪ …	**Financial (borrower):** ▪ Assets; ▪ Obligations; ▪ Monthly income; ▪ Monthly expenses; ▪ …
Credit history (borrower): ▪ In the current bank; ▪ In other banks; ▪ Credit bureau data; ▪ …	**Social indicators (borrower):** ▪ Work experience; ▪ Time of residence at the current address; ▪ Marital status; ▪ …

The specificity of indicator values in the scoring card is formed as a dynamic database of client interactions.

In the classic setting, the task of scoring analysis is formalized as the task of classifying new customers based on information about existing clients.

Let the set of bank customers be given

$$\{Z_i\},\ i=1\ldots n\,. \tag{1}$$

Each client is characterized by a p-dimensional vector of heterogeneouss features.

$$X_i=[x_{i1},\ldots x_{ip}]^{\mathrm{T}}\,. \tag{2}$$

It is known that each client Y_i belongs to one of two creditworthiness classes $k\le 2$:

$$Y=\begin{cases} y=1-\text{the client is creditworthy}, \\ y=0-\textit{the client is not creditworthy}. \end{cases} \tag{3}$$

New clients are characterized by a sample: $\{W_j\},\ j=1\ldots m$.

A sample of clients with known creditworthiness class serves as the training set — $\{Z_i^N\},\ i=1\ldots(n-N)$.

It is necessary to implement a scoring algorithm that classifies new clients $\{W_j\}$, $j = 1 \ldots m$, based on their feature vectors $X_i = [x_{i1}, \ldots x_{ip}]^{\mathrm{T}}$.

The specified scheme corresponds to the strategy of learning with a teacher, however, in the practice of scoring analysis, the presence of a training sample is a rather rare phenomenon. In this case, forming a training set becomes a separate, rather complex task.

Additionally, discrete indicators of the scoring card may not be informative when considered individually. This necessitates calculating secondary indicators and comparing them with other clients.

Therefore, the article considers a modified formulation of the problem of scoring analysis.

Let a set of bank clients (1) be given. Each client is characterized by a p-dimensional vector of heterogeneous features (2) $X_i = [x_{i1}, \ldots x_{ip}]^{\mathrm{T}}$. It is necessary to classify each bank client Y_i into two classes (3).

The multi-criteria scoring method is proposed to solve the classification problem formalized in this way. This decision is based on the following considerations. Scoring analysis is essentially aimed at building a digital twin of the banking institution's team, which forms the requirements for the ideal client. This enables the analysis of alternatives for binary classification of clients based on a multitude of factors, comparing the ideal image with the real client — evaluating the degree of closeness between them. This process is accompanied by considerations/doubts/analysis of many factors—often following the "best–worst" scheme. It is multicriteria scoring that allows incorporating the decision maker's considerations in transforming static indicators into dynamic requirements of criteria.

In addition, the scoring model must meet a number of technical requirements [1–5]:

1. High adequacy in dividing borrowers into two categories from the perspective of credit issuance: "positive" and "negative";
2. The scoring point is a measure of the probability of the borrower belonging to the "positive" or "negative" class;
3. The scoring model should form the average rating of "negative" borrowers significantly lower than the average rating of "positive" ones;
4. There should be a ranking of borrowers within the rating of "positive" decisions;
5. There is a cut-off point when it is unprofitable for the bank to issue loans to borrowers below a certain scoring point;
6. The scoring model should ensure detection of fraud.

Multi-criteria scoring implements the given list of requirements and has a number of unique advantages, which will be proved with a computational example [8].

Based on the structure of the scoring card (Table 1) and setting extremum requirements for its indicators, we generally obtain a system of criteria for the categories of the scoring card:

credit product (bank)

$$P = [p_i \to \text{extreme}]^{\mathrm{T}}, \; i = 1 \ldots k_p, \tag{4}$$

financial (borrower)

$$F=[f_i \to \text{extreme}]^{\mathrm{T}},\ i=1\ldots k_f\,,$$

credit history (borrower)

$$K=[k_i \to \text{extreme}]^{\mathrm{T}},\ i=1\ldots k_k\,,$$

social (borrower)

$$S=[s_i \to \text{extreme}]^{\mathrm{T}},\ i=1\ldots k_s\,,$$

extended vector of criteria

$$W=[w_i = p_i, f_i, k_i, s_i \to \text{extreme}]^{\mathrm{T}},\quad i=1\ldots k_p + k_f + k_k + k_s\,.$$

The directions of the extremum (*extreme=min, max*) of each indicator of the scoring card are unique for each banking institution. This effectively reflects the bank's understanding of the image of the ideal client and considers the logic of mental deliberations "best-case scenario – worst-case scenario".

Analysis of the content and practical significance of indicators in the scoring card suggests a conflicting nature of criteria for the "ideal" borrower. Therefore, we have a multicriteria optimization problem in scoring.

The decision-making model is formed by aggregating/integrating partial criteria vectors (4) into a generalized/integrated assessment score using convolution through a non-linear trade-off scheme [8].

Compared to other aggregation schemes of partial criteria [9], convolution has a number of proven advantages [8]. The convolution uses a non-linear trade-off scheme, which allows obtaining a Pareto-optimal solution with low computational costs. The optimization problem is solved under constraints, ensuring unimodality of the generalized criterion function and guaranteeing a unique solution in any case. Convolution enables the use of a minimax approach, focusing on maximizing the dominant partial criterion of optimality. Weight coefficients of partial criteria allow consideration of subjective factors in dominating their influence on scoring results.

The convolution criterion for discretely given partial optimality criteria has the form [8]:

$$Y(\varphi_0)=\sum_{l=1}^{b}\gamma_{0l}(1-\varphi_{0l})^{-1} \to \min\,, \tag{5}$$

where $l=1\ldots b$ — the number of partial optimality criteria included in the convolution; γ_{0l} — normalized weight coefficient; φ_{0l} — normative partial criterion.

The values of weight coefficients are assigned within a unified rating scale and normalized according to the expression:

$$\gamma_{0l}=\frac{\gamma_l}{\sum_{l=1}^{b}\gamma_l}, \tag{6}$$

where γ_l is the current (non-normalized) value of the weight coefficient.

The normalization of partial criteria aims to bring them to a single scale of change (0...1) and to the direction of minimization. Therefore, partial criteria that are minimized and those that are maximized are normalized separately.

Normalization can be implemented, for example, relative to the maximum (minimum) values characterizing the change in partial optimality criteria by expressions

$$\varphi_{0l}^{\min} = \frac{\varphi_l^{\min}}{\max \varphi_l^{\min} + \Delta}, \ \varphi_{0l}^{\max} = \frac{\min \varphi_l^{\max} - \Delta}{\varphi_l^{\max}}, \tag{7}$$

where $\max \varphi_l^{\min}$, $\min \varphi_l^{\max}$ — maximum and minimum values of the minimizing and maximizing criteria in the interval of their consideration; Δ — the reserve coefficient, which varies between 0.1 and 0.3 and ensures the elimination of the operation of division by zero for normalization of values $\max \varphi_l^{\min}$, $\min \varphi_l^{\max}$.

Convolution (5) can be presented in matrix form

$$Y(\varphi_0) = G\Phi^T,$$
$$G = [\gamma_1, \gamma_2, \gamma_3, \ldots, \gamma_l], \ \ l = 1 \ldots b, \tag{8}$$
$$\Phi = [(1-\varphi_{01})^{-1}, (1-\varphi_{02})^{-1}, (1-\varphi_{03})^{-1}, \ldots, (1-\varphi_{0l})^{-1}]^T, \ \ l = 1 \ldots b.$$

To form a multi-criteria mathematical model of bank scoring, we formalize the appearance of the scoring card in the accepted notation (4) — see table 2.

Table 2. The scoring card — formalized structure

№	P				F				K				S			
	p_1	p_2	…	$p_{i=k_p}$	f_1	f_2	…	$f_{i=k_f}$	k_1	k_2	…	$k_{i=k_k}$	s_1	s_2	…	$s_{i=k_s}$
	w_1	w_2	…	w_{k_p}	w_{1+k_p}	w_{2+k_p}	…	$w_{k_p+k_f}$	…	$w_{k_p+k_f+k_k+k_s}$						
1	$w_{1(1)}$	$w_{1(2)}$	…	$w_{1(k_p)}$	$w_{1(1+k_p)}$	$w_{1(2+k_p)}$	…	$w_{1(k_p+k_f)}$	…	$w_{1(k_p+k_f+k_k+k_s)}$						
2	$w_{2(1)}$	$w_{2(2)}$	…	$w_{2(k_p)}$	$w_{2(1+k_p)}$	$w_{2(2+k_p)}$	…	$w_{2(k_p+k_f)}$	…	$w_{2(k_p+k_f+k_k+k_s)}$						
…	…	…	…	…	…	…	…	…	…	…						
ν	$w_{\nu(1)}$	$w_{\nu(2)}$	…	$w_{\nu(k_p)}$	$w_{\nu(1+k_p)}$	$w_{\nu(2+k_p)}$	…	$w_{\nu(k_p+k_f)}$	…	$w_{\nu(k_p+k_f+k_k+k_s)}$						

Taking into account the given designations, the generalized assessment of the ν-th borrower according to the vector of criterion requirements (4) for the scoring card of Table 1 in accordance with the convolution (5) is determined by the expression:

by the extended vector of criteria in scalar form:

$$Y_\nu(w_0) = \sum_{l=1}^{k_p+k_f+k_k+k_s} \gamma_{\nu(0l)} (1 - w_{\nu(0l)})^{-1} \to \min, \tag{9}$$

by the extended vector of criteria in matrix form:

$$Y_\nu(w_0) = G_\nu \Phi_\nu^{\mathrm{T}},$$
$$G_\nu = [\gamma_{(1)}, \gamma_{(2)}, \gamma_{(3)}, \ldots, \gamma_{(l)}], \ \ l = 1 \ldots k_p + k_f + k_k + k_s, \tag{10}$$
$$\Phi_\nu = [(1-\mathrm{w}_{\nu(01)})^{-1}, (1-\mathrm{w}_{\nu(02)})^{-1}, (1-\mathrm{w}_{\nu(03)})^{-1}, \ldots, (1-\mathrm{w}_{\nu(0l)})^{-1}]^T,$$
$$l = 1 \ldots k_p + k_f + k_k + k_s.$$

Normalization of weight coefficients, as well as partial criteria included in (9), (10), is implemented according to expressions (6), (7), taking into account the direction of the extremum.

To account a significant number of criteria in the generalized multicriteria assessment, it is advisable to use the technology of nested convolutions. This approach also allows regulating the influence of groups of scoring card indicators on the assessment result. This is implemented by sequentially (within the four groups of partial criteria (4)) reduction of partial criteria to the generalized by group and to the integrated efficiency criterion in scalar form:

$$P_{\text{v}}(p_0) = \sum_{l=1}^{k_p} \gamma_{\text{v}(0l)} (1 - p_{\text{v}(0l)})^{-1} \to \min,$$

$$F_{\text{v}}(f_0) = \sum_{l=1}^{k_f} \gamma_{\text{v}(0l)} (1 - f_{\text{v}(0l)})^{-1} \to \min,$$

$$K_{\text{v}}(k_0) = \sum_{l=1}^{k_k} \gamma_{\text{v}(0l)} (1 - k_{\text{v}(0l)})^{-1} \to \min,$$

$$S_{\text{v}}(s_0) = \sum_{l=1}^{s} \gamma_{\text{v}(0l)} (1 - s_{\text{v}(0l)})^{-1} \to \min,$$

$$Y_{\text{v}}(w_0) = \gamma^{P}_{\text{v}(0l)} (1 - P_{\text{v}0}(p_0))^{-1} + \gamma^{F}_{\text{v}(0l)} (1 - F_{\text{v}0}(f_0))^{-1} +$$

$$+ \gamma^{K}_{\text{v}(0l)} (1 - K_{\text{v}0}(k_0))^{-1} + \gamma^{S}_{\text{v}(0l)} (1 - S_{\text{v}0}(s_0))^{-1} \to \min, \tag{11}$$

$$P_{\text{v}0}(p_0) = [\sum_{i=1}^{k_p} (1 - [\max p_{\text{v}(0l)} - \Delta])^{-1}]^{-1},$$

$$F(p_0) = [\sum_{i=1}^{k_p} (1 - [\max f_{\text{v}(0l)} - \Delta])^{-1}]^{-1},$$

$$K_{\text{v}0}(p_0) = [\sum_{i=1}^{k_p} (1 - [\max k_{\text{v}(0l)} - \Delta])^{-1}]^{-1},$$

$$S_{\text{v}0}(p_0) = [\sum_{i=1}^{k_p} (1 - [\max s_{\text{v}(0l)} - \Delta])^{-1}]^{-1}. \tag{12}$$

Normalization (12) is performed relative to the worst assessment — the maximum value of the normalized indicator, which characterizes the partial criterion of the scoring card.

Similarly, matrices are formed and generalized group criteria ratings, which are part of the matrix model of multicriteria scoring (10), are normalized.

$$Y_{\text{v}}(w_0) = \gamma^{P}_{\text{v}(0l)} (1 - G_{pv0} \Phi^{T}_{pv0})^{-1} + \gamma^{F}_{\text{v}(0l)} (1 - G_{fv0} \Phi^{T}_{fv0})^{-1}$$

$$+ \gamma^{K}_{\text{v}(0l)} (1 - G_{kv0} \Phi^{T}_{kv0})^{-1} + \gamma^{S}_{\text{v}(0l)} (1 - G_{sv0} \Phi^{T}_{sv0})^{-1} \to \min,$$

$$G_{pv0}, G_{fv0}, G_{kv0}, G_{sv0} = \text{normalization}[\gamma_{\text{v}(0l)}], \quad l = 1 \ldots k_p, k_f, k_k, k_s$$

are the same in structure, but may have different values

$$Y_{\text{v}}(w_0) = \gamma^{P}_{\text{v}(0l)}(1 - G_{pv0}\Phi^{\text{T}}_{pv0})^{-1} + \gamma^{F}_{\text{v}(0l)}(1 - G_{fv0}\Phi^{\text{T}}_{fv0})^{-1}$$

$$+ \gamma^{K}_{\text{v}(0l)}(1 - G_{kv0}\Phi^{\text{T}}_{kv0})^{-1} + \gamma^{S}_{\text{v}(0l)}(1 - G_{sv0}\Phi^{\text{T}}_{sv0})^{-1} \to \min\,, \qquad (13)$$

$$G_{pv0}, G_{fv0}, G_{kv0}, G_{sv0} = \text{normalization}[\gamma_{\text{v}(0l)}]\,,$$

$l = 1 \ldots k_p, k_f, k_k, k_s$ are the same in structure, but may have different values

$$\Phi_{pv0} = \text{normalization}[(1 - p_{\text{v}(0l)})^{-1}]^{\text{T}}\,,\ \ l = 1 \ldots k_p\,,$$

$$\Phi_{fv0} = \text{normalization}[(1 - f_{\text{v}(0l)})^{-1}]^{\text{T}}\,,\ \ l = 1 \ldots k_f\,,$$

$$\Phi_{kv0} = \text{normalization}[(1 - k_{\text{v}(0l)})^{-1}]^{\text{T}}\,,\ \ l = 1 \ldots k_k\,, \qquad (14)$$

$$\Phi_{sv0} = \text{normalization}[(1 - s_{\text{v}(0l)})^{-1}]^{\text{T}}\,,\ \ l = 1 \ldots k_s\,.$$

The interpretation of the obtained result involves bringing the value of the generalized assessment (9), (10), or in the form (11), (13) to a unified scale, for example, from 0 (the worst rating) to 1 (the best rating). This is achieved by normalizing the generalized score to the abstract worst customer score according to the expression

$$I_0 = 1 - \frac{I}{\max I}\,, \quad \max I = \sum_{i=1}^{5}(1 - [\max F_i - \Delta])^{-1}\,, \qquad (15)$$

where $\max F_l$ — the worst possible value of the partial indicator; Δ — the reserve coefficient, which ensures the avoidance of incorrect operations during normalization.

The obtained numerical assessment can be converted to the linguistic category of the client's solvency according to the fundamental assessment scale of the Table 2.

Table 3. Fundamental rating scale

Integrated performance assessment I_0	Linguistic category of efficiency
1,0 – 0,7	High
0,7 – 0,5	Good
0,5 – 0,4	Satisfactory
0,4 – 0,2	Low
0,2 and less	Unsatisfactory

The numerical evaluation of the normalized generalized indicator (15) (see the left column of Table 3) is proportional to the probability of the client returning the loan. That is, it characterizes the risk of providing a credit loan.

Thus, expressions (9)–(15) form *a multi-criteria mathematical model of credit scoring*. The differences, advantages and features of the model are as follows. The model allows you to consider the indicators of the scoring card in terms of infological connections: factors, indicators, criteria, which contributes to increasing the adequacy of the ideal client profile of a banking institution. The model ensures obtaining a generalized client assessment as a solution to an optimization problem using a minimax approach to image requirements. The obtained solution is Pareto-optimal. Subjective priorities of scoring card indicators can be taken into account in client assessment by adjusting both partial and group criteriaweights. The model is structurally open to adding scoring card indicators. The proposed model does not require a priori data on loan issuance/refusal to clients, that is, it implements an unsupervised learning scheme. Thus, the proposed model can be a primary superstructure, acting as a highly accurate binary classifier to deep learning methods based on artificial neural networks. Undoubtedly, artificial neural networks are designed and capable of accumulating large segments of labeled data, and the proposed multi-criteria model cannot compete with these advantages. But in the context of unsupervised learning, the multi-criteria model has better potential properties than the existing approaches in the essence of the formalization of the classification task. Research presented below has proven the model's capability to detect fraud. The model fully meets the requirements for bank scoring models, which will also be proved by a computational example.

II. *The methodology of multi-criteria credit scoring* determines the sequence of actions for for performing calculations and obtaining the resulting assessment, including the following stages:

1. Establishing a set of indicators from the scoring card (Table 1) in the form of (4).
2. Normalizing criteria (4) using expressions (6), (7).
3. Formulating the generalized client assessment, expressions (9)–(14).
4. Interpreting the generalized assessment with normalization (15) and in accordance with Table 3.

III. *The technology of multicriteria credit scoring* involves practical aspects of implementing the synthesized model (9)–(15) and the methodology of its application to the architecture of the software system and to a specific script-based implementation.

The technological processes *of multi-criteria credit scoring can be represented by the structural diagram in Fig. 1, which implements the architecture of the software script of credit scoring.*

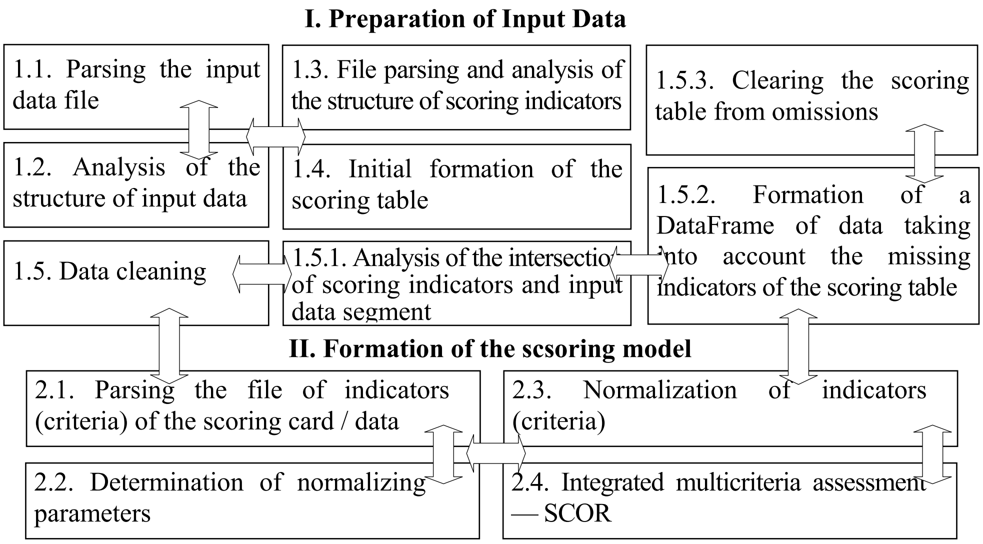


Fig. 1. Structural diagram of the software implementation of the mathematical model

Technological processes are divided into two blocks: preparation of input data and formation of the scoring model. The input data includes two files: scoring card indicators in "sample_data.xlsx" and indicator descriptions in "data_description.xlsx" (Fig. 2). In total, there are 121 scoring card indicators and 500 records of potential bank clients.

Application	loan_amoun	loan_days	applied_at	gender	birth_date	Marital	children	education	fact_addr	fact_addr_start	has_immov
1	3000	30	01.02.2021 0:21	1	03.02.1995	2	1	5	4	14.06.2015	0
2	1000	7	01.02.2021 0:24	1	19.01.1984	2	2	4	0	NULL	0
3	1000	3	01.02.2021 0:36	2	02.08.1994	2	1	4	2	17.05.2018	0
4	1600	30	01.02.2021 0:34	1	12.11.1992	1	1	3	3	25.11.1992	0
5	2500	18	01.02.2021 23:22	2	22.10.1997	1	1	5	3	01.01.1999	1
6	1000	30	01.02.2021 0:38	2	12.06.1996	2	1	3	3	10.12.2009	0
7	1300	30	01.02.2021 0:40	1	05.04.1993	1	1	5	1	26.09.2012	1
8	2000	14	01.02.2021 1:00	1	17.09.1984	2	2	5	3	19.09.2012	0
9	2500	12	01.02.2021 0:47	1	27.09.1986	1	1	5	1	12.01.2015	0
10	1000	7	01.02.2021 0:50	2	15.06.1991	1	1	6	4	01.09.2014	0
11	2000	30	01.02.2021 0:52	2	21.08.1981	5	3	3	1	12.03.1981	0
12	1400	3	01.02.2021 0:50	2	04.04.1994	4	1	4	2	29.07.2010	0
13	2500	30	01.02.2021 1:19	1	27.06.1991	2	2	3	2	27.09.2007	0
14	2000	30	01.02.2021 1:04	1	11.03.1998	1	1	4	2	08.05.1998	0
15	1500	30	01.02.2021 2:17	1	09.08.1996	1	1	2	1	27.10.2002	0
16	1000	30	01.02.2021 1:40	1	25.02.1995	1	1	5	3	25.02.1995	0

a — Scoring card sample_data.xlsx

Field_in_data	Description_of_information	Filling	At_the_time	Place_of_definition	Note
id	Application ID	Necessarily	Known	Determined by the application registration system at the time of submission	
loan_amount	Loan amount in the application	Necessarily	Known	Specified by the borrower	
loan_days	Loan repayment term in the application	Necessarily	Known	Specified by the borrower	
applied_at	Date and time of application acceptance	Necessarily	Known	Determined at the time of application	Local time on the user's device or application time?
language				Determined by the browser at the time of application	Always UA
purpose_id	ID of the purpose of loan use	Necessarily	Known		Always 999-999, the client enters the destination itself, previously it was possible to choose from a list
purpose_other	The purpose of using the loan	Necessarily	Known	Specified by the borrower	
ip	IP of the client from which the request came	Necessarily	Known	Determined by the browser at the time of application	
loan_status_id	Not used		Not known	Assigned by the application processing system	
decision_status_id	The status of the decision on the application		???		
decision_application_	ID in the decision-making system		???		
decision_flow_id	ID the credit policy process		???		
decision_is_exported	Export check box		???		
decision_exported_at	Export date		???	**fields are related to decision-making**	
product_id	ID of the product issued to the customer	Necessarily	???	Parameters related to this product	

b — Indicators of the scoring card data description.xlsx

Fig. 2. Structure of input data

The implementation of credit scoring technology in script form, according to the developed model and its application methodology, is carried out using the Python programming language with libraries such as NumPy and Pandas. **The original project with the code is available at** https://github.com/Pysarchuk-O/scoring.git .

Over the input data files, a series of preparatory stages are implemented, dictated by the specifics and details of the data: parsing of data files and converting them to the pandas – DataFrame format (blocks 1.1, 1.3); analyzing the structure of the input data by size, data types, and presence of missing values (block 1.2); initial formation of the scoring card and analysis of its structure for compliance with the fields of Table 1 (block 1.4); cleaning the input data from gaps using the rejection strategy in cases of a significant number of them and the impossibility of recovery based on the essence of the indicator (registration address, residence, place of work, etc.); selection of values for the scoring card based on the intersection of input data (Fig. 2, *a*) and indicators (Fig. 2, *b*) with further control of preservation of the of Table 1's structure (block 1.5). After these stages, which are classic, a content analysis of indicators is implemented (also within Block 1.5's functionality). This involves selecting objective-subjective indicators that purely characterize the borrower and are not secondary, as designated by the system (product status, product profile ID, etc.). To specify the data in Fig. 2, these are indicators from the scoring card marking fields (see Fig. 2, *b*) "borrower indication" and "parameters related to the issued product". Analysis showed that the data in Fig. 2 do not have solutions for issuing the product, there is no data structure of the "training pair" type. Thus, fields related to "decision-making" (Fig. 2, *b*) are absent in the scoring card (Fig. 2, *a*). Therefore, we are dealing with data oriented towards implementing an unsupervised machine learning model.

The result of the data preparation stage is a scoring card in general form, as shown in Fig. 2, *a*, which contains a list of indicators presented in Fig. 3. Analysis of the scoring card structure according to Fig. 2 and comparison with Table 1 confirms the correspondence and presence of the main categories for scoring.

	A	B	C	D	E	F	G
1		**Field_in_data**	**Description_of_information**	**Filling**	**At_the_time**	**Place_of_definition**	**Note**
2	**0**	loan_amount	Loan amount in the application	Necessarily	Known	Specified by the borrower	
3	**1**	loan_days	Loan repayment term in the application	Necessarily	Known	Specified by the borrower	
4	**2**	gender_id	Gender of the borrower	Necessarily	Known	Specified by the borrower	ID decryption is required
5	**3**	birth_date	Borrower's date of birth	Necessarily	Known	Specified by the borrower	Sometimes it is filled incorrectly
6	**4**	children_count_ic	The number of children of the borrower	Necessarily	Known	Specified by the borrower	ID decryption is required
7	**5**	education_id	The borrower's education level	Necessarily	Known	Specified by the borrower	ID decryption is required
8	**6**	fact_addr_owner	Residential address: property type		Known	Specified by the borrower	ID decryption is required
9	**7**	has_immovables	Owns real estate	Necessarily	Known	Specified by the borrower	

Fig. 3. Scoring card indicators after data cleaning and balancing d_segment_data_ description_cleaning.xlsx

Further, a series of stages for script deployment and application of the proposed multicriteria credit scoring model is implemented. This is implemented in accordance with the formulated methodology and is reflected in the structural diagram in Fig. 1 by blocks 2.1–2.4.

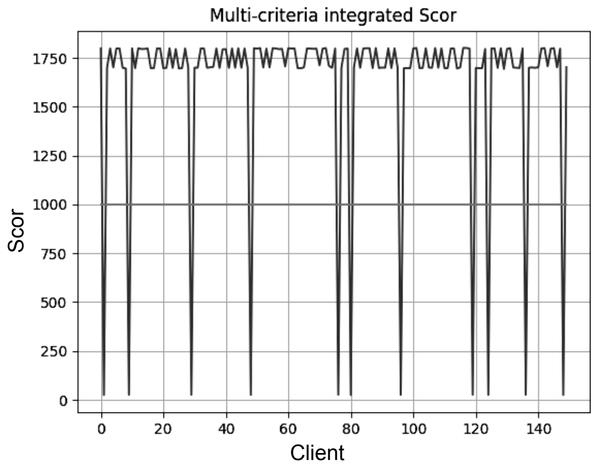
The initial step to initiate the model's operation involves parsing the indicators (criteria) file of the scoring card (block 2.1). The criteria requirements for the scoring card indicators are formulated based on the values remaining after cleaning and balancing (Fig. 3). These criteria should reflect the ideal client profile and are unique to each banking institution. It is possible, for example, to establish a system of criteria requirements shown in Fig. 4. They are used for the calculation example.

	A	B	C	D	E	F	G
1		**Field_in_data**	**Description_of_information**	**Minimax**	**At_the_time**	**Place_of_definition**	**Note**
2	**0**	loan_amount	Loan amount in the application	min	Known	Specified by the borrower	
3	**1**	loan_days	Loan repayment term in the application	min	Known	Specified by the borrower	
4	**2**	gender_id	Gender of the borrower	max	Known	Specified by the borrower	ID decryption is requi
5	**3**	birth_date	Borrower's date of birth		Known	Specified by the borrower	Sometimes it is filled incorrectly
6	**4**	children_count_id	The number of children of the borrower	min	Known	Specified by the borrower	ID decryption is requi
7	**5**	education_id	The borrower's education level	max	Known	Specified by the borrower	ID decryption is requi
8	**6**	fact_addr_owner_type_	Residential address: property type		Known	Specified by the borrower	ID decryption is requi
9	**7**	has_immovables	Owns real estate	max	Known	Specified by the borrower	
10	**8**	has_movables	Owns a vehicle	max	Known	Specified by the borrower	
11	**9**	employment_type_id	Place of work: type of employment	max	Known	Specified by the borrower	ID decryption is requi
12	**10**	organization_type_id	Place of work: type of organization	max	Known	Specified by the borrower	ID decryption is requi
13	**11**	organization_branch_id	Place of work: branch of the organizatio	max	Known	Specified by the borrower	ID decryption is requi
14	**12**	employees_count_id	Place of work: number of employees of	max	Known	Specified by the borrower	ID decryption is requi

Fig. 4. Criterion requirements for scoring card indicators d_segment_data_description_ cleaning_minimax.xlsx

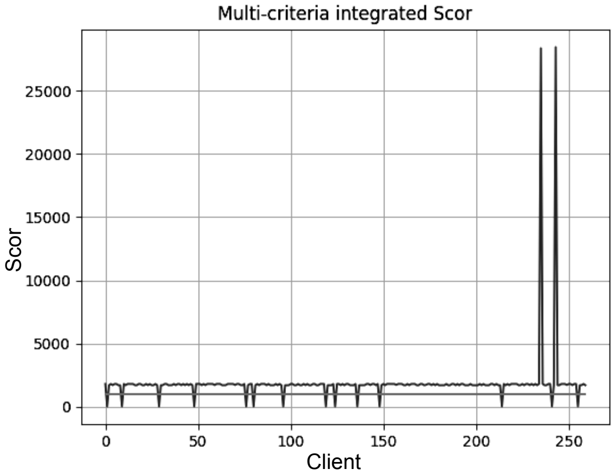
Further, the selection of minimax indicators indicated in Fig. 4 from the scoring map of Fig. 1 and the formation of an integrated multicriteria assessment (scor-py) according to model (9) is implemented. The calculation results are presented in the graphs of Fig. 5.

The graphs in Fig. 5 show the unnormalized scoring values (scor axis) for each client included in the scoring card in Fig. 2, *a* (client axis). Fig. 5, *a* presents the scoring evaluation for 150 clients without abnormally high scores. Fig. 5, *b* and 5, *c* show the scoring evaluations for 250 and all 500 clients with abnormally high score values. The red line characterizes the empirical value of dividing customers into creditworthy and non-creditworthy. The results shown in Fig. 5, *a* demonstrate a strictly binary classification of clients into the mentioned categories. Differences in the scores of creditworthy clients reflect their internal distribution, proportional to the risk characteristics of granting credit. Abnormally high scoring values presented in Fig. 5, *b* and 5, *c* indicate possible fraud in the data or incorrectly filled fields in the scoring card. Fraud analysis involves reversing the process of analyzing the criteria vector for each indicator individually and in the multicriteria scoring complex.

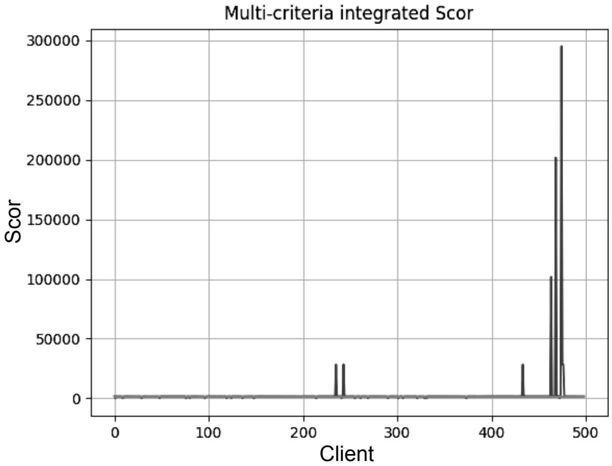


a — The scoring evaluation for 150 clients without abnormally high scores

Fig. 5. Client Creditworthiness Assessment — scor (*Beginning*)



b — The scoring evaluations for 250 clients with abnormally high score values



c — The scoring evaluations for 500 clients with abnormally high score values

Fig. 5. Client Creditworthiness Assessment — scor (*Continued*)

Comparison of the above results of calculations obtained in accordance with the proposed approach was carried out with discriminant analysis on a similar segment of data. The results of the comparison in more than 80% did not contradict each other, which indicates the effectiveness of the proposed approach.

CONCLUSIONS

In the research, a multi-criteria mathematical model of credit scoring for Data Science tasks was developed. The model provides the formation of the ideal client profile in the format of criterial requirements. The integrated assessment is formed using convolution according to a nonlinear trade-off scheme and represents the solution to an optimization problem, belonging to Pareto optimal solutions. The model operates on data which corresponds to the principles of unsupervised learning. It can be applied independently or serve as a foundational layer for deep learning methods. The practical application of the model is facilitated by its methodological and technological representations. A computational example of applying the model to real big data proved its effectiveness for the decision-making stage of customer lending and fraud detection.

REFERENCES

1. Mohamed Hani AbdElHamid, Mohamed Tawfik ElMasry, *Machine learning approach for credit score analysis: a case study of predicting mortgage loan defaults*. NOVA Information Management School Instituto Superior de Estatística e Gestão de Informação Universidade Nova de Lisboa, 2019, 60p. Available: https://run.unl.pt/bitstream/10362/62427/ 4/TEGI0439.pdf
2. Maria Fernandez Vidal, Fernando Barbon, *Credit scoring in financial inclusion. How to use advanced analytics to build credit-scoring models that increase access.* Washington: Consultative Group to Assist the Poor: World Bank, 2019, 52 p. Available: https://www.findevgateway.org/guide-toolkit/2019/07/credit-scoring-financial-inclusion
3. "Credit scoring approaches guidelines," *The World Bank Group*, 2019, 64 p. Available: https://thedocs.worldbank.org/en/doc/935891585869698451-0130022020/ original/ CREDITSCORINGAPPROACHESGUIDELINESFINALWEB.pdf
4. Rory P. Bunker, M. Asif Naeem, Wenjun Zhang, "Improving a Credit Scoring Model by Incorporating Bank Statement Derived Features," *Preprint submitted to Expert Systems with Applications*, 2017. Available: https://www.researchgate.net/ publication/309606871_Improving_a_Credit_Scoring_Model_by_Incorporating_Bank_Statement_Derived_Features
5. Engku Muhammad Nazri E.A. Bakar, "Credit scoring models: techniques and issues," *Journal of Advanced Research in Business and Management Studies*, vol. 7, issue 2, pp. 29–41, 2017. Available: https://www.researchgate.net/publication/323306120_Credit_scoring_models_techniques_and_issues
6. Gerrit Muller, *System Modeling and Analysis: a Practical Approach*. University of South-Eastern Norway-NISE, 2023, 128 p. Available: https://www.gaudisite.nl/ SystemModelingAndAnalysisBook.pdf
7. Nolberto Munier, *Mathematical Modelling of Decision Problems. Using the SIMUS Method for Complex Scenarios*. Springer, 2021, 196 p.
8. Albert Voronin, *Multi-Criteria Decision Making for the Management of Complex Systems*. IGI Global, 2017, 220 p.
9. Goran Cirovi´c, Dragan Pamuˇcar, *Multiple-Criteria Decision Making*. MDPI, 2022, 313 p.

Received 10.07.2024

INFORMATION ON THE ARTICLE

Oleksii O. Pysarchuk, ORCID: 0000-0001-5271-0248, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine, e-mail: PlatinumPA2212@gmail.com

Maria D. Vasylieva, ORCID: 0000-0003-2217-643X, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine, e-mail: mdvasilieva@gmail.com

Danylo R. Baran, ORCID: 0000-0002-3251-8897, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine, e-mail: danil.baran15@gmail.com

Illya O. Pysarchuk, ORCID: 0000-0003-4343-0142, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine, e-mail: flimka134@gmail.com

БАГАТОКРИТЕРІАЛЬНА МАТЕМАТИЧНА МОДЕЛЬ КРЕДИТНОГО СКОРИНГУ В ЗАДАЧАХ DATA SCIENCE / О.О. Писарчук, М.Д. Васильєва, Д.Р. Баран, І.О. Писарчук

Анотація. Запропоновано багатокритеріальну оптимізаційну математичну модель кредитного скорингу. Модель отримано з використанням нелінійної схеми компромісів для вирішення задач багатокритеріальної оптимізації, що дозволяє побудувати Парето-оптимальне рішення. Запропонований підхід формує інтегровану оцінку кредитоспроможності позичальника на основі структурованого набору показників, що відображають фінансовий, кредитний та соціальний профіль клієнтів. Модель призначено для використання в інтелектуальних CRM та ERP системах, що працюють з великими даними, і не потребує розмічених навчальних вибірок, що робить її придатною для задач навчання без учителя. Вона також може слугувати базовим рівнем для подальшого аналізу з використанням методів глибинного навчання. Описано методологічні кроки впровадження моделі, від нормалізації показників до прийняття остаточних рішень. Технологічна реалізація демонструє ефективність моделі в автоматизованому прийнятті рішень щодо кредитування і виявленні шахрайства.

Ключові слова: Data Science, Big Data, SCORIG машинне навчання, прийняття рішень, багатокритеріальні математичні моделі, інтелектуальні CRM, ERP системи.