Customer Classification and Decision Making in the Digital Economy based on Scoring Models

HENNADII MAZUR[1], NATALIA BURKINA[2], YURII POPOVSKYI[2], NADIIA VASYLENKO[1], VOLODYMYR ZAIACHKOVSKYI[1], RUSLAN LAVROV[3], SERHII KOZLOVSKYI[4]
[1]Department of Management and Administration,
PHEE «Vinnytsia Academy of Continuing Education»,
Vinnytsia, UKRAINE
[2]Department of Marketing and Business Analytics, Vasyl' Stus Donetsk National University,
Vinnytsia, UKRAINE
[3]Department of Economics, Finance and Accounting,
PHEI «European University», Kyiv, UKRAINE
[4]Department of Entrepreneurship, Corporate and Spatial Economics,
Vasyl' Stus Donetsk National University, Vinnytsia, UKRAINE

Abstract - The article presents the way of applying cluster models to customer classification and managerial decision on retaining the available clients and acquiring new ones. The objective of the research is to find out the relevant techniques for building scoring models in different fields. The main research was testing the hypothesis: if the number of point models is approximated in different spheres of activity, then the proposed methods will be universal. To check this hypothesis the vector method of k-nearest neighbors support was applied for decision making in the digital economy based on scoring models. In order to realize the principle of customer classification and revealing the client categories with risk of quitting, the client's classification model was created. Moreover, a risk issue was shown in the example of fraud dynamic. Different fraud categories were studied to define their features. On the basis of the model building results, the authors proposed some recommendations on decision making in risk situations. The model shows how to retain existing clients and how to share client base through the client groups and how to deal with risks of losing clients.

Key-Words: - Modelling, decision making, algorithms, scoring models, customer classification, digital economy, cluster analysis.

Received: October 15, 2022. Revised: February 29, 2023. Accepted: March 21, 2023. Published: April 6, 2023.

1 Introduction

Development and widespread use of computer technologies have caused significant changes in economies. Many activities increasingly rely on the Internet. In many spheres (for example, insurance, health, science, agriculture), information is digitized, [1]. All they allow to accumulate and create large volumes of data (Big Data), processing of which becomes an additional vector of socio-economic, technical, scientific analysis, establish new logical patterns and to make managerial decisions based on them, [2].

The development and visual effectiveness of digital technologies have a significant impact on the economy and society. The electronic toolkit leads to radical changes in people's lives and is one of the priorities for leading countries in economy, including the United States, Germany, China, Japan etc. The creation of new technologies including the ones for business brings about global changes associated with the emergence of new digital infrastructures, the rapid development of digital communications and the improvement of computer technology.

The integration of these technologies into the economic and socio-political life of society testifies to the formation of a new system of the world economy – a digital economy, [3]. IT technologies are constantly improving and getting cheaper. New technologies emerge to better interact with customers and promptly respond to various changes in business relations. Lagging behind at the technology market means bearing risks and becoming uncompetitive or even ejected from the market. What is scoring and what features does it have? Making a bank profit directly depends on the quality of the loan portfolio. The smaller the financial risks, the greater the likelihood of a quick return of borrowed funds with an additional profit from the payment of interest. That's why, considering loan applications, the bank carries out a thorough check up of potential customers for possible financial risks. For each client, the entrepreneur estimates the prospect of retaining the person as a client for a maximum possible time. For this purpose, the risk of possible client loss is again assessed. Prior to launching a product to the market, an entrepreneur estimates the possible buyers and assesses all the risks.

In all these cases economists use scoring models that help reduce risks of the economic activity, find, accumulate and support customers, optimize the production and supply of new products, as well as increase profits. Scoring is a heuristic way of developing ratings and classifying different objects into groups. It assumes that people with similar social indicators behave the same way. Today scoring is used in banking, marketing, insurance and many other spheres of business activity. Literally, "scoring" means counting of scores. This article shows what kind of points modern analysts consider and what for they use them.

2 Literature Review

The term "digital economy" is very popular nowadays. But there is no single definition of it. The dDigital economy is the creation, distribution and use of digital technologies and related products and services. Digital technologies are technologies for the collection, storage, processing, retrieval, transmission and submission of data electronically, [4]. Richard Heeks in his research on Information and Communication Technologies for Development, points out that these technologies create new opportunities in the digital sphere: an entrepreneur or a company can optionally use the digital system in their activities, [3]. This process may include datafication (implementation of storage technologies for large arrays), digitization (conversion of all parts of the value chain from the analogue to digital format), virtualization (physical fractioning of processes), and generativity (use of data and technology in new, different from the original, assignments by reprogramming and recombination), [3]. Thus, as generalized by R. Bucht, [5], the term "digital economy" refers exclusively to the events that are currently underway and the unfinished transformation of all sectors of the economy due to the digitization of information by using computer technology. It is explained by the fact, that the informational technologies enable payments on line, concluding online business agreements.

The digital economy is the most important issue of scientific debate and especially its legal aspects. Nowadays, the digital economy is an unrestricted way of doing business and requires the establishment of sound state control mechanisms for the legal activity of economic entities. The study of the regulation of digital business interactions is a strategic task of national and interstate policy aimed at ensuring the security of the entire modern world. Undoubtedly, each state strives to ensure legal economic interaction with its citizens. It is common knowledge, that thanks to the Internet, the operation of online stores and online shopping is not always legally supported and protected. This requires definition of concepts and conditions at the regulatory level in different countries of the world. In most cases, misunderstandings and disputes arise in unforeseen situations. For example, buying real estate with the possibility of electronic signature, which in turn is a risk of e-commerce.

Certainly, often controversial situations that arise in the digital environment can be resolved through current legislation. The difficulty is usually caused by the fact that many legal provisions, of course, do not explicitly provide for their application to Internet relations. Under such circumstances, an interpretation of the legal provisions is required. Studies of Cherdantsev V., Kobelev P. point out that the digital economy encompasses a complex of electronic business operations and e-commerce, as well as the corresponding infrastructure. The process of informatization influences the regulation of production, assets circulation and e-commerce, incorporating the global information network for communication, [6]. F. Mishko, K.Vasilyeva, V. Popov in their research "Trends in the Legal Regulation of Civil and Competitive Relations in the Digital Economy" say that implementation of digital agreements requires separation of concepts of digital offer and digital acceptance and expanding the list of legal objects by including the terms "information" and "digital assets", [7]. Digital economy is evolving not only as a tool of purchasing online or concluding business agreements, but also as a tool for market research and decision-making. Using smart technologies of economic processes interaction, it would be advisable to develop the direction of smart contract models.

A. Vashkevich in his work "Smart Contracts: What and Why" in detail describes that a significant part of the norms can be algorithmized and the regulation become automated. Smart contracts have become known along with blockchain and cryptocurrency technology. Now they are part of reality. But smart contracts are much wider than algorithms that translate tokens. They will change the legislative reality, lawyers' work and business life, [8]. Access to data in digital economy became a key factor in product development and innovations, and data collection and their use by the third parties causes the controversial issues concerning protection of the competition and the rights of businessmen. An important factor of successful international business in the field of digital economy is ensuring honest competition. Rapid growth of innovations and use of new technologies within digital economy sometimes advances traditional models of regulation therefore state policy can consider not fully the growing competition in various industries.

In scientific research "Feature of legal regulation of digital intellectual economy" K.M. Belikov says, "to guarantee the open markets, innovations, quality and efficiency and also freedom of choice for consumers, the effective competition has to be protected from restrictions", [9].

Instruments of protection of the competition in the conditions of digital economy embrace the following: ban on the conclusion of anti-competitive agreements; the ban on abuse of a dominant position in the market; control of merges for prevention of domination in the market and prevention of creation of essential obstacles for the effective competition, [10].

Different approaches to legal regulation in the sphere of digital economy come down to the same conclusion, that it is necessary to provide such a legal regime which would enable fee development of innovations and prevent possible risks. Since there exists a risk of disability to precisely predict the development of innovations in digital economy, the new legislation must be flexible and take into account a great number of data, [11]. Around the world the global financial institutions face various problems among which digital fraud is not the least. Rapid development of communications and information technologies and their active application in the sphere of financial services causes various types of fraud, and billions of dollar losses. For the purpose of prevention of fraud in the sphere of business by the company of the Lab neurodat it is developed technology of assessment of emotions of people on the basis of the set parameters. Specially under this task developed the Emotion Miner platform which continues work and allows to analyze video. Collected data formed the basis of methods of training of neuronets in recognition of human emotions. Algorithms pay attention to a voice (tone height, a timbre, loudness, pauses in language), emotional coloring and semantics of the text, a mimicry the person, speed and the direction of movements of a body and position of separate extremities, heart rate on the basis of changes of skin color, breath on the movement of a thorax and also a sex, age of the person and presence at it on a face of points, moustaches, beards.

The result of work received multimodal architecture which at the same time can analyze audio, video, gestures and physiological parameters. Developments are planned to be used in branches of business, advertising, spheres of safety and medicine, [12], [13].

Initially, scoring was designed to automate the process of deciding on a loan. Prior to the introduction of scoring, the decision who is to be issued the loan to was made by a credit expert. Basing on his experience, he decided on the client's creditworthiness. In the 1940s, the implementation of scoring systems began. In 1941, David Durant published the first credit scoring research to evaluate the role of various factors in the forecasting system. After the end of World War II, the demand for credit products increased sharply, and it became clear that traditional decision-making methods were performing poorly for large numbers of customers, [14]. The explosion in demand for loans, driven in part by the implementation of credit cards, motivated lenders to introduce automated systems for deciding on lending. The development of computer technology made it possible to process large amounts of financial data. In 1956, FICO was established to develop consumer loans. In the 60's, the implementation of computer technology in the scoring area began. In 1963, it was proposed to use discriminant data analysis for credit scoring. In 1975 with the adoption of the "US Equal Credit Opportunity Act I", the scoring was finally recognized, [15]. An important step in the development of credit scoring was the emergence of behavior scoring in the early 90's. Its purpose is to predict payments to existing customers.

Recently, the development of scoring systems has been driven by external regulation. As part of the capital adequacy requirements for banks following the entry into force of the second Basel Committee for Banking Supervision 2001, institutions should closely monitor the risks associated with their loan portfolios, [16]. Credit scoring methods allow to do this. In Christina Bolton's dissertation "Logistic regressions and their application in credit scoring", (2009) the concept of credit scoring for banking in South Africa is considered. It points out the methods of constructing a scoring model with emphasis on the logistic regression method, [17]. The thesis of Matthias Kremlin "Adaptive models and their application in credit scoring", (2011) analyzes methods of constructing predictive models in the conditions of drift and data retention. A new method for building scoring models based on the decision tree method is presented. It's used to estimate drift in two sets of real financial data, [18].

3 Problem Statement and Hypotheses of Research

3.1 Research Objectives

The purpose of the research is theoretical justification and development of cluster models working for creating customer classification and making managerial decision. It might help retain available clients and find new ones.

According to the aim of the research, the following research tasks have been formulated:

– to analyze different approaches to defining the digital economics and digital environment;
– to demonstrate features of the legal regulation in the digital economics;
– to consider scoring models in different spheres of economics and law;
– to define their types and advantages using in different fields of economics and law;
– to adopt k-nearest neighbors support vector method for implementing the principle of customer classification and revealing the client categories with risk of leaving the company;
– to discuss different ways of determining risk groups of clients;
– to show the effect of applying mathematical models in the example of the company.
– to propose some recommendations on retaining the existing clients and acquiring new ones.

3.2 Purpose of the Study

This paper is aimed to study the relevant techniques for building scoring models in the spheres of economy and legislature. Some basic scoring models, their types and advantages for different economic domains were discussed. For practical feasibility, the k-nearest neighbor vs support vector method was used in order to implement the principle of customer classification and to reveal the clients intending to leave the company.

3.3 The Research Hypothesis

The main hypothesis of the research was to the quantity of scoring models in different fields. Probably, among the variety of scoring models there is a set of more effective decision-making ones. The main attention was paid for finding risk groups of clients. The issue under study is: "What scoring models categories are the most effective for decision making while working with risk clients according to comparative analysis". Thus, thanks to the results of the working model a set of the most relevant models was found and some proposals were made on retaining the existing clients through creating the client group portfolios. Scoring is a whole customer distribution system based on statistics. It is an important assistant in determining the potential solvency or the future activity of the client as well as the reliable helper of prompt assessment, which is widely used in the economic and legal sector today. The main goal of traditional scoring is to classify bank customers into two categories - "good" and "bad", depending on the lender's decision on the further actions with this client, [19]. A "bad" client, for example, can be defined as a client with a low empirical probability of loan repayment. To make a decision, the financial institution issuing a loan to the borrower uses a system for calculating points. Data processing for decision making is assigned to algorithms using scoring. Test tasks are being developed, as it were, to sort out risk zones and automatically calculate the borrower's potential solvency. If decision-making algorithms are in the risk zone, the client may be offered a smaller amount or other conditions, [20]. The decision will be made based on many factors. The introduction of artificial intelligence, which laid down the conditions of economic processes and developed a mechanism for calculating the criteria, allows to develop the assessment system.

Scoring is a complex mathematical algorithm that can draw conclusions based on processed data, analyze social factors on an existing client base in a few years. For example, a scoring program can process data on defaulters or debtors over the past some years and identify typical social, age, or behavioral factors, [21]. Based on these data, the evaluation will be adjusted and, when analyzing the next clients, the program will consider these new factors. Obviously, in banking databases, you can use algorithms that will look for similar characteristics in new loan applications relative to past similar contacts. It should be noted that scoring is not an ideal financial risk analysis program but it helps to quickly and accurately make managerial decisions when working with big data, [22].

4 Methodology

According the scoring tasks, all scoring models are divided into three categories, demonstrated in Figure 1.

Application scoring – scoring credit rating of an individual. If, after entering all the answers in the program, the loan officer/mortgage banker claims that the scoring has been completed, this means the completion of the bulk of the analytical verification. Next, the person's loan application goes to the security service, where bank experts check the client according to several criteria. Conducting a scoring assessment can eliminate the human factor – both specialist's bias towards a client and overly loyal attitude as well as intentional concealment of some factors that indicate an increased financial risk for the bank, [23], [24]. The financial scoring algorithm is quite complicated and considers many factors when setting a general assessment of financial risks. Each bank has its own algorithm for verifying customer solvency and discipline regarding loan repayments.

Credit scoring is an automatic scoring system for a borrower. Each client of the bank fills up a questionnaire containing detailed personal information. Each of his/her characteristics has points of certain value. After checking the reliability of these data and summing up the scores, a decision is made on the solvency of the potential borrower and, based on this value, on the issuance or non-granting of a loan. The value of the "passing" score depends on the loan product.

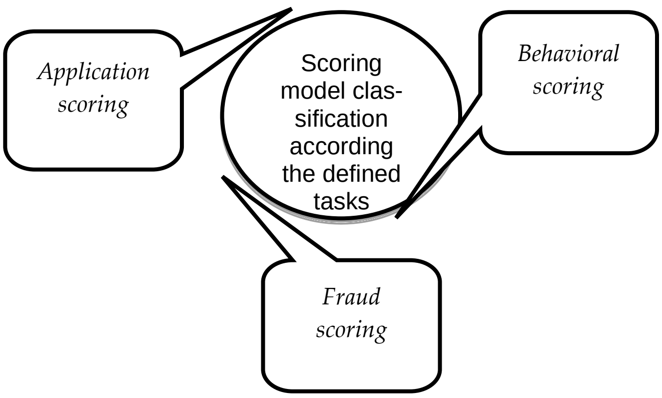

Fig. 1: Scoring model classification according to the defined tasks
Source: compiled by the authors

Scoring cards consist of hundreds of positions constantly updated and changed. They are created based on processing large amounts of data on credit precedents: repaid and outstanding loans. For example, statistics show that women are more disciplined in financial matters and therefore have a higher score. The factors of a person's residence in the area, as well as his employment in an industry, have their own values. This value depends on the current economic depression of the region and the growth or decline of production. Persons with conviction records, administrative offenses, non-payment of fines or alimony have a significantly lower scoring. In addition to points, there are so-called stop- and go-factors - circumstances that clearly block the consideration of the borrower's application or, on the contrary, immediately give it a "go". For example, the first is the applicant's age (too young or too old), the second – work in a prestigious international company or in a company that has been the bank's client for many years.

Fraud scoring. This type of scoring is a complex system for detecting any inconsistencies or matches that are also detected through cross-checks. Its goal is to identify anything that might arouse suspicion, [25], [26]. When the loan application arrives, the client's personal data are first checked for authenticity. They are checked through various databases that banks purchase from law enforcement agencies and credit bureaus. Bankers can also use publicly available data. For example - a database of invalid, stolen or lost identity documents. If the authentication was endorsed, the rules of cross-checks for identifying suspicious situations begin to work. A lot of factors are analyzed and compared: phone numbers, customer addresses, addresses of bank branches, names of tank managers who arrange loans, age of borrowers and others. For example, the system will respond if a new loan application contains a business phone number, which in several previous ones was indicated as home, because it considers this to be suspicious. It will report that the applicant is registered at the same address as the person listed by the bank in the "blacklist". Scoring considers that one of the family members with suspicious or negative past inclines relatives to fraud. If the system has found something suspicious, it issues two scenarios. The first is automatic failure. It is issued if there are clear signs of fraud. For example, the application contains a passport, which is listed in stolen items, or a contact phone number, which is on the bank's blacklist. The second – the application is submitted to the risk managers of the bank for “manual” verification, [27]. It happens if a circumstance is found without obvious criminal signs, but it requires explanation. For example, two loan applications have the same address of residence and home phone number. Perhaps these are people who live in a civil marriage, or maybe it's a dummy phone and address. In this case, the verifier calls one of the clients and finds out whether these people know each other, clarifies some parameters in order to understand the credibility of the client’s words. Risk management constantly monitors changes in the quality of the bank's loan product portfolios and develops new audit rules. Each bank maintains its blacklist of customers, which is constantly updated. Security services cooperate with each other and with law enforcement agencies. Such scoring helps identify fraudsters by a variety of signs that they often don’t even know about. However, it cannot foresee all situations.

Another classification of scoring models bases on the mathematical methods for building these models, [28]. Among the statistical methods are popular discriminant analysis, linear regression, logistic regression and decision tree. Other methods originate from mathematics: mathematical programming, neural networks, genetic algorithms and expert systems. Let’s analyze the most common methods represented in Figure 2.

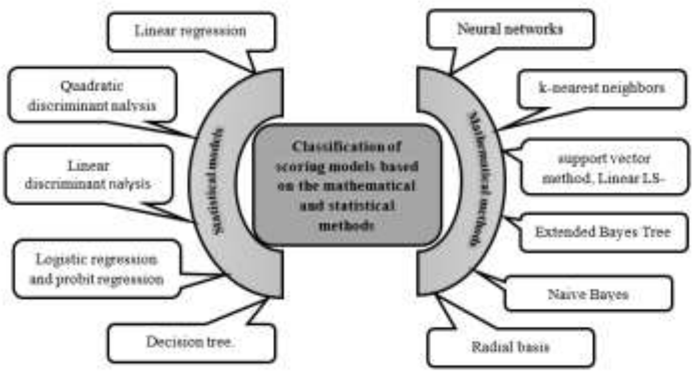

Fig. 2: Classification of scoring models based on the mathematical and statistical methods
Source: Compiled by the authors

Linear discriminant analysis (LDA). Linear discriminant analysis is a method for classifying objects into predefined categories. Its main idea is to find a linear combination of explanatory variables that would best categorize objects. By separation, it is best understood as one that ensures the maximum distance between the average of these categories. The score is calculated as a linear function of the client’s attributes values:

$$Z = \beta^T x = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k, \qquad (1)$$

where $x = (x_1, \ldots, x_k)$ – customer attribute values, $\beta = (\beta_1, \ldots, \beta_k)$ – model parameters that maximize the ratio,

$$M = \frac{\beta^T (m_G - m_B)}{\sqrt{\beta^T \Sigma \beta}} \qquad (2)$$

m_G, m_B is the vector of means for good and bad customers, Σ is the general covariance matrix.

The linear discriminant method involves the fulfillment of two conditions:

– the covariance matrices of independent variables for both groups must coincide.
– independent variables should be distributed normally.

The main advantage of this method is the possibility to use it even in case of normality violation.

Quadratic discriminant analysis (QDA) is a nonlinear generalization of the LDA. It’s a method that does not use the assumption of homogeneity of the covariance matrix. As a decision rule, a quadratic function (3) is applied:

$$d_k = -0.5(x - \mu_k)^T C_k^{-1} (x - \mu_k) - 0.5 \ln |C_k| + \ln \pi_k \qquad (3)$$

where |Ck| is the determinant of the covariance matrix of the k-th class, C_k^{-1} its inverse matrix; π_k is the priori probability of observing objects of the k-th class. The test object also belongs to the class with the maximum value d_k.

A quadratic discriminant analysis is very effective when the dividing surface between the classes has a pronounced nonlinear character (for example, a paraboloid or an ellipsoid in the 3D case). However, it retains most of the LDA shortcomings: it uses the assumption that the distribution is normal and does not work when covariance matrices degenerate (for example, with many variables). Another disadvantage of QDA is disability to “explain” the results because of the equation of the separating hypersurface is expressed implicitly.

In marketing, discriminant analysis is often used to identify factors that differentiate between different types of customers and/or products based on surveys or other forms of data collection. The use of discriminant analysis in marketing is usually described by the following steps:

1. Formulating a task and collection data. Identification of the most significant features by which buyers evaluate a product in this category. Use quantitative marketing research methods (such as surveys) to collect data from a group of potential buyers judging by their preferences for the characteristics of the goods. Data for various products are encoded and entered into a statistical system, such as R, SPSS, or SAS.

2. Estimate the coefficients of the discriminant function and determine the statistical significance and consistency. Choose the appropriate discriminant analysis method. Direct methods evaluate discriminant function with simultaneously evaluation of all attributes. A step-by-step method introduces the features sequentially. Two-class methods should be used when the dependent variable has only two states. The multiple discriminant method is used when the dependent quantity has three or more qualitative states. SPSS uses Wilk's Lambda or F-stat in SAS to test significance. The most common method of evaluating solvency is to divide the available data into estimates and verification or deferred data. Evaluation data is used to construct the discriminant function. The deferred data is used to construct a classification matrix, which indicates the number of correctly and incorrectly classified objects, [29].

3. Drawing a two-dimensional picture, determination the dimensions, interpretation the results. A statistical program marks the results. In two-dimensional space, each object is displayed. The distance between products characterizes the degree of difference between them. Dimensions must be determined by the researcher. This requires subjective judgment and is often a daunting task.

Linear regression. A linear regression method is the simplest scoring method. In the case of two categories, it is equivalent to the linear discriminant analysis method and expresses the dependence of one variable (dependent) on the other (independent). In general, it's represented by formula (4):

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \varepsilon \tag{4}$$

where Y – dependent variable; X_i – explaining independent variables; β_i – unknown regression coefficients that are found by the least squares method; ε – error.

It requires the following assumption: the relation between the dependent and independent variables must be linear; errors should be independent and distributed normally, [30].

Logistic regression and probit regression. The logistic regression model is binary model. It allows to model and to forecast simple categorical data. The logistic regression model is defined as follows:

$$\log\left(\frac{p}{1-p}\right) = \beta^T x = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \tag{5}$$

where p is the estimate of the probability that the client is “bad”, β is the vector of unknown regression parameters, which is calculated as maximizing the likelihood ratio.

The logistic regression model is based on the logarithm function. In turn, probit regression is based on a normal distribution and is defined as follows:

$$N^{-1}(p) = \beta^T x = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \tag{6}$$

where $N(x) = \frac{1}{2\pi}\int_{-\infty}^{x} e^{-\frac{y^2}{2}} dy$; the vector β is calculated like the logistic regression model.

Since logistic regression and probit regression use similar distribution shapes, the results of using these models are also similar. Logistic regression is highly preferred, since it enables simpler than in probit regression calculations and more tools to work with it. Due to its binary nature, logistic regression is preferable to linear regression in use for building scoring models. In practice, it was found that the difference in the accuracy of the predicted results is insignificant. However, there is a predominance of logistic regression in scoring systems, [31].

Neural networks. Artificial neural networks are simulations of neural networks found in nature. Neural networks consist of layers which, in turn, consist of nodes. There are 3 types of layers in networks: input, hidden, output. Customer attributes, such as gender, age, etc., form the input layer. The output y_k for the k-th node with m inputs is represented as follows:

$$y_k = \varphi(v_k) = \varphi\left(\sum_{j=0}^{m} \omega_j x_j\right) = \varphi(\omega^T x) \tag{7}$$

where $\varphi(x)$ is the activation function, x is the input data vector, ω is the weight vector which indicates the strength of the connection between nodes.

Despite the possibility of achieving high accuracy of the forecast, it is impossible to understand the reasons why a decision was made. It's the main disadvantage of the use of neural networks for scoring models development.

Method k nearest neighbors. Nonparametric method for classifying objects is based on a metric that determines the similarity between the data. Initially, training data is divided into classes, then the evaluated data are entered and the similarity between the entered and training data is determined. Based on the metric, k nearest neighbors are selected. The new element belongs to the class with the most of its neighbors. The number of neighbors k is determined by a compromise between compensation and dispersion. The smaller the class, the less k is chosen. Moreover, the result for the large k will not necessarily be better. One of the advantages of this method is the easy possibility to add new data without changing the model, [32]. The nonparametric nature of this method allows to work with irrationalities in risk functions in the attribute space. The absence of a formal method for choosing k and the impossibility of a probabilistic interpretation of the result, are the main disadvantages of the method. These difficulties can be solved using the Bayesian approximation method.

Comparison of various methods. A series of comparative studies have been conducted for scoring methods. The ranking criteria were the percentage of classification errors and the ROC curve. Eight data sets have been studied (Table 1).

The Table 1 shows that Neural Networks, Support Vector Method and Logistic Regression were the best in the studied eight data sets, [33]. There is no optimal scoring model for any situation. The choice of model depends on the data and the purpose of creating the model. In addition, the best rating method will not necessarily be the best in this situation.

5 Statistical Analyses

5.1 Reliability and Validity

Aiming to achieve high reliability and validity of the research, in the calculations, [34], it was applied the set of scoring models as well as a probability statistical model. The first model using the data of real gym company help to build the system of decision making for risk clients. And the second one describing the fraud processes was based on the official open data statistics. It shows the level of different types of frauds and the dynamic of their changes. It helps to account this factor working with unknown clients.

Table 1. Comparison of various methods

Method	Average rating
Neural networks	3.2
Support Vector Method	3.7
Logistic Regression	4.3
Linear discriminant analysis	5.3
Linear LS-SVM	5.5
Extended Bayes Tree	5.6
Naive Bayes Classifier	7.8
Radial basis functions	9.1
k-nearest neighbors (k =100)	9.5
Linear SVM	10.1
Quadratic discriminant analysis	10.8
Decision tree	10.8
Linear programming	11.9
Decision tree	13.7
k-nearest neighbors (k = 10)	14.1

Source: Compiled by the authors

Data collection

To develop Customers Classification Scoring Model, it was considered sample of the gym clients, which was consist of the next features:

Age – age of client;

Income – average client's month income, thousand $;

Children – the number of children or grandchildren under the age of 15;

Sex – male (1) or female (0);

Education – school (1); Bachelor's degree (2); Master's degree (3); compulsory school or secondary school certificates (4);

Visit Count – the number of people visiting a gym during the last month;

Is Client – if a person remains the client of the gym (1) or he leaves this gym (0).

For the fraud statistical analyses, official open data were used, [13].

5.2 Data Analysis and Results

The mathematical model of Behavioral scoring for customers classification according k-nearest neighbors vs support vector method is demonstrated in Figure 3. The model is created at the Stat Soft Statistica Enterprise 10.0 with the help of module k-means clustering after normalization of the entering sample.

Hennadii Mazur, Natalia Burkina,
Yurii Popovskyi, Nadiia Vasylenko,
Volodymyr Zaiachkovskyi, Ruslan Lavrov,
Serhii Kozlovskyi

Figure 3 shows what categories of clients have the biggest probability to leave this gym next month. So, there are two risk groups at this gym. The riskiest group is represented by the Cluster 1 and the next risk group is shown in the Cluster 5.

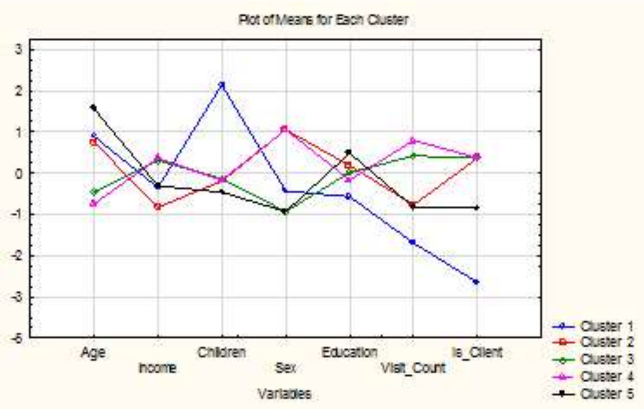

Fig. 3: Customer Classification Scoring Model
Source: Compiled by the authors

These clusters consist of clients with the following characteristics (Figure. 4). In this way, some risk client categories were determined, among which there was a group characterized as the group with middle income, aged above 50 with little grandchildren having no high school diploma, [35], [36].

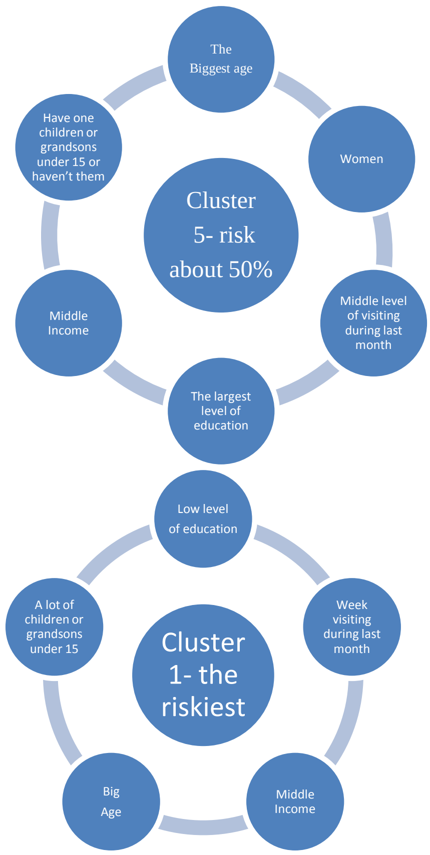

Fig. 4: Risk clusters client's characteristics
Source: Compiled by the authors

The next set of probably risk clients is the group of women past 60 with middle income, having some diplomas and/or degrees without little grandchildren or probably only with one grandson.

Finally, the set of elderly gym visitors was singled out. The necessity of turning attention to elderly people was highlighted. It was pointed out that the considered company needs at least two advertising programs: for elderly people and their grandchildren and for elderly clients without little grandchildren, [37]. The next analysis was conducted according to fraud statistics. There are a lot of methods of fraud detection. Among them are: identity validation potential risks associated with the borrower's individual characteristics; phone and address check to validate borrower's information; Income and Employment Analysis; variety of automatic systems like MERS (Mortgage Electronic Registration System) or NFPB (National Fraud Protection Database) and others. The aim of this research was to understand the way frauds occur and what trends in fraud emerge.

Let's consider the official statistics of fraud by financial products, demonstrated in Figure 5.

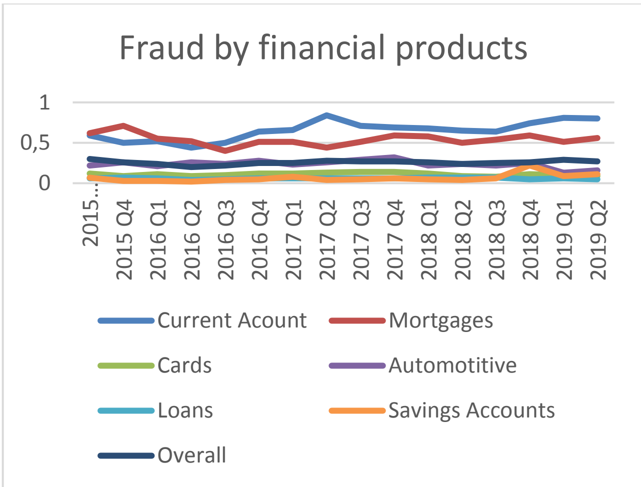

Fig. 5: Fraud by financial products
Source: Compiled by the authors

Given graphs demonstrate that the level of fraud is not very high. Different kinds of fraud have different variations, but all of them are less than 1%, [38], [39]. The riskiest activity is related with Current Accounts. Moreover, this kind of frauds has the highest growth dynamic. The less risky operations are connected with Saving Accounts. However, it also shows a rise dynamic during the last year.

Table 2 shows Mean and Standard Deviations in different kinds of frauds.

Mean value demonstrates the two riskiest operations: Current Account and Mortgages and operations almost without risk: Loans and Saving Account. Standard Deviation shows the same trend – the biggest variation for the first two operations and the least – for the last two kinds of fraud. Moreover, the deviation for all operations except Current Account and Mortgages is less than error of calculation.

Table 2. Mean and Standard Deviations different kinds of fraud

Variable	Mean and Standard Deviations Case wise deletion of MD N=16	
Current Account	0,650625	0,117840
Mortgages	0,540000	0,071926
Cards	0,109375	0,018786
Automotive	0,236875	0,046435
Loans	0,059375	0,008539
Savings Accounts	0,065000	0,047610

Source: Compiled by the authors

In order to define dependence between all given kinds of fraud a correlation analysis was conducted. Correlation matrix shown in Table 3 consists of the Pearson correlation coefficients and demonstrates the level of dependence between different frauds.

Table 3. Correlation matrix for different kinds of fraud

Variable	Current Account	Mortgages	Cards	Automotive	Loans	Saving Accounts	Overall
Current Account	1,00	-0,17	0,33	-0,26	0,13	0,49	0,68
Mortgages	-0,17	1,00	-0,08	0,04	0,30	0,18	0,32
Cards	0,33	-0,08	1,00	0,51	0,04	0,01	0,42
Automotive	-0,26	0,04	0,51	1,00	0,01	-0,22	-0,23
Loans	0,13	0,30	0,04	0,01	1,00	-0,27	0,37
Savings Accounts	0,49	0,18	0,01	-0,22	-0,27	1,00	0,32
Overall	0,68	0,32	0,42	-0,23	0,37	0,32	1,00

Source: Compiled by the authors

Correlation matrix helps to understand relations between different types of fraud. Thus, it highlights that frauds on Current Accounts have the most influence on the Overall frauds with the correlation level *r*=0,68. And the Saving Accounts fraud influence on the Current Account fraud with the *r*=0,49. This fact explains the trend of growth of both types of financial fraud in the last year. So, despite of the minor level of Saving Account fraud, considering its rising trend, there is a growth risk of the biggest level of fraud – Current Account.

In order to consider the character of fraud's changing it's interesting to research Frequency Distribution for each kind of fraud. They are demonstrated in Figures 6-11. To identify their features, a comparative analysis of their graphs was carried out. According to the given graphs, all of them have normal distribution, but their own parameters and properties differ, [40], [41].

Figure 6 demonstrates Frequency Distribution for Current Account.

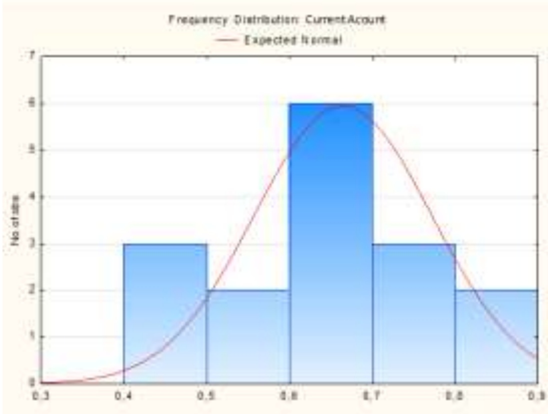

Fig. 6: Frequency Distribution for Current Account
Source: Compiled by the authors

Usually, normal curves are symmetric about the mean μ. But practically it has some variations in its parameters. So, the Frequency Distribution of Current Account is left-skewed or negatively-skewed distribution because it has a little longer left tail in the negative direction in the number line. The mean in this direction is also to the left of the peak. Moreover, it has the mean to the left of the median. Thus, frauds in Current Account are extremely important to research because they have the biggest value, and more than half considered years have the percent more than mean. Moreover, it has the growth trend. To conclude, it's the riskiest kind of financial operations considering frauds (Figure 7).

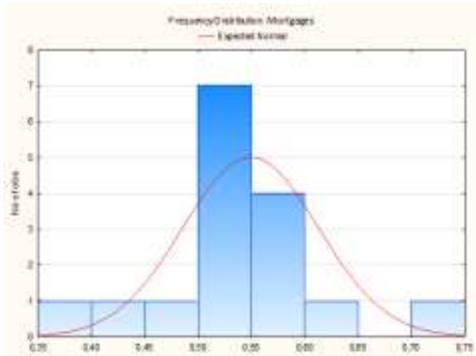

Fig. 7: Frequency Distribution for Mortgages
Source: Compiled by the authors

Frequency Distribution for Mortgages is almost symmetric Normal Distribution. But it has a positive value of Kurtosis, which tells that it has heavy-tails or a lot of data in its tails. It means that the mean value doesn't characterize Mortgages fraud for all considering years. It was a random high peak value that maybe requires additional research. It's more typical to have a low value for this distribution. So, Mortgages fraud does not present danger, but sometimes there is some unknown activity that requires more thorough consideration (Figure 8).

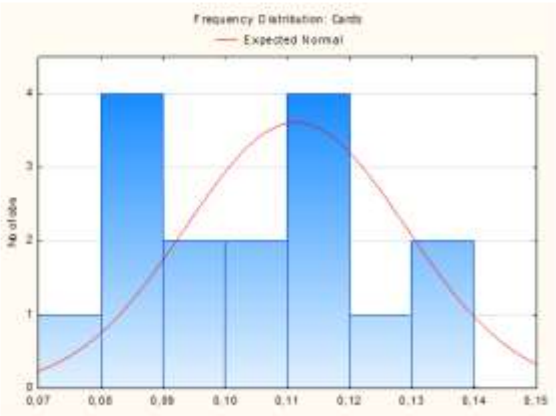

Fig. 8: Frequency Distribution for Cards
Source: Compiled by the authors

In case of Cards operation, it's interesting to note the bimodal character of Normal Distribution. It is characterized by two peaks and it means that there is no clear value of cards fraud. So, the risk of frauds in these operations is rather high, because it's difficult to forecast the level of such frauds at each time period. Despite of the stochastic character of this distribution, the level of danger in this situation isn't so high due to non-critical level of mean according to statistical values (Figure 9).

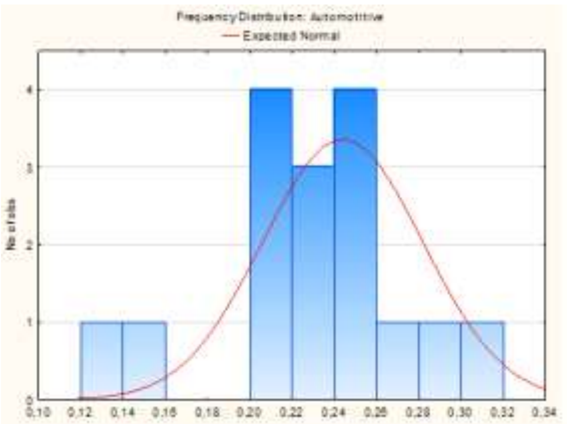

Fig. 9: Frequency Distribution for Automotive
Source: Compiled by the authors

Distribution of Automotive frauds partly resembles Frequency Distribution for Current Account, but with a lower high peak. So, it's not so dangerous as Current Account fraud, but it's recommended to pay attention to it. Moreover, according to the statistical analysis, demonstrated in Figure 4, this type of frauds occupies the third place among all kinds of all financial frauds (Figure 10).

Saving Accounts have right-skewed distribution with a long right tail. Right-skewed distributions are also called positive-skew distributions, because there is a long tail in the positive direction on the number line. The mean is also to the right of the peak. Because this histogram's tail has the biggest positive skew to the right, Saving Accounts frauds have light-tails or little data in their tails, especially in their right tail. This fact once more proves that this kind of frauds is the least risky (Figure 11).

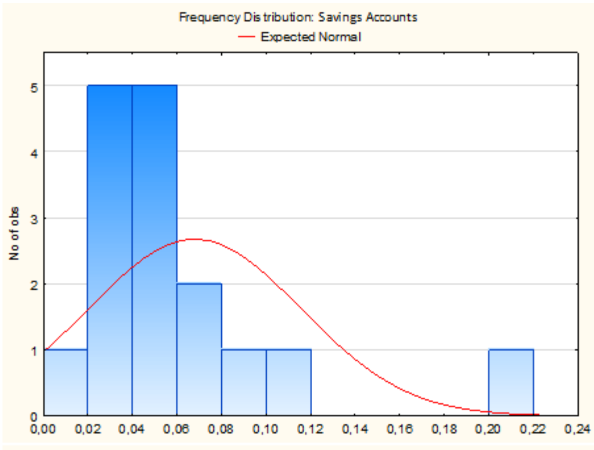

Fig. 10: Frequency Distribution for Saving Accounts
Source: Compiled by the authors

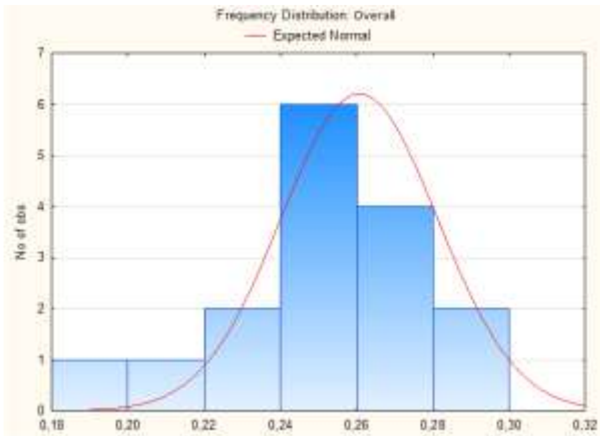

Fig. 11: Frequency Distribution for Overall
Source: Compiled by the authors

The overall distribution of financial frauds has similar form to Current Account Distribution. And it's not surprising, because the current accounts according the correlation analysis have rather high influence on the overall frauds.

Thus, the statistical and visual analysis of distributions give the opportunity to find out some features of fraud character. The most and the less risky financial operations were determined and hypotheses about their further behavior were put forward.

6 Discussion

6.1 Key Findings and Results

The results of the study are summarized in the following statements:

– methods of building scoring models in different fields of economy and law have been analyzed and investigated in order to create classifications of client groups.
– Classifications of client groups for financial managerial decisions have been developed and theoretically substantiated.
– Evaluation models aiding managers in determining potential solvency or future activity of the client, as well as in rapid assessment of financial characteristics of clients have been developed and considered.
– Scoring models, their types and advantages of use in different spheres of economy are considered, as well as methods of reference vectors of k-nearest neighbors are specified for implementation of the principle of classification of clients and identification of clients with risk of leaving the company.
– The risk groups of clients are determined based on the use of scoring models and are practically proven by the mathematical model in the example of the company.
– Based on the results of the model, the way of retaining existing clients and sharing the client base by client groups portfolios was proposed.
– Factors of influence of new information products and technologies on the modern economic market are considered. A comparative analysis of the services market using new technologies and rapid interaction with customers is provided.
– Risks in economic structures, which depend on timely response to various changes in business relations and lag from the technology market, resulting in the risk of non-competitiveness or displacement from the market have been considered.

6.2 Prospects for Further Research

The relevance of the results of the study is proven by the fact that the development of information technologies leads to applying new tools in business management. Modern Internet technologies are developing very rapidly and financial frauds a new chance and way. Therefore, business and financial structures should make steps in search for new methods of early financial risk prevention, [34], [42]. Information technology combined with mathematical statistical methods enables decision-making algorithms to be constructed well in advance of loan application processing prior to financial transactions.

Given the availability of complete statistical information, further research should turn to decision-making technologies in the socio-economic domain incorporating the Internet options; development of computer models, risk and consumption loss algorithms, investment into analytical technologies.

7 Conclusion

The research shows that the better scoring system is developed, the more objective it is and the more correctly and quickly it will evaluate the bank risks preventing it from possible losses. That is why each enterprise develops its own scoring model according to its target client group and keeps it secret. The way of misleading this model is to know how to answer specific questions in the questionnaire. And this is the main reason why enterprises almost never report to the customers about the reasons for the refusal. Scoring has its own strong and weak points. It helps to identify potential default clients and fraudsters by eliminating the risks of issuing a loan to an unreliable client or refusing to a reliable one. This research has described the most widely used methods for constructing scoring models. Currently, scoring is widely used all over the world and has proven to be an effective decision-making tool in the digital economy. In many spheres of economy, expert assessment services have been replaced by more reliable and credible scoring systems. However, despite its wide use and prolific works of foreign scientists, scoring is not adequately studied in the Ukrainian economics. Scoring has great potential for use but is still a "black box" for people using it. Scoring systems are undoubtedly worth further studies and improvements.

References:

[1] Dahlman, C., Mealy S., Wermelinger M. (2016). *Harnessing the Digital Economy for Developing Countries*. Paris: OECD. Retrieved from http://www.oecd-ilibrary.org/docserver/download/4adffb24-en.pdf

[2] Haltiwanger J., Jarmin R.S. (2000). Measuring the Digital Economy. Understanding the Digital Economy (E. Brynjolfsson, B. Kahin (eds)). Cambridge: MIT Press, MA, pp. 13–33.

[3] Bucht, R., Hicks, R. (2018). Definition, concept and measurement of the digital economy. *Bulletin of international organizations*, 13, 143-172.

[4] Kozlovskyi, S., Nikolenko, L., Peresada, O., Pokhyliuk, O., Yatchuk, O., Bolgarova, N., Kulhanik, O. (2020). Estimation level of public welfare on the basis of methods of intellectual analysis. *Global Journal of Environmental Science and Management*, 6(3), 355-372. Retrieved from http://dx.doi.org/10.22034/gjesm.2020.03.06

[5] Bucht R., Hicks R. Definition (2019). *Concept and measurement digital economy*. Retrieved from Retrieved from https://iorj.hse.ru/data/2018/08/30/1154589879/Бухт Хикс Определение Концепция и измерение цифровой экономики.pdf

[6] Cherdantsev, V., Kobelev, P. (2010). Formation of a single information space. *Agrarian Gazette of the Ural*, 11-1 (77), 102-103.

[7] Myshko, F., Vasileva, K., Popov, V., Strelnikov, I. (2019). Trends in legal regulation of civil and competitive relations in the sphere of digital economy. *Journal of Economic Security*, 1, 155-159.

[8] Vashkevich, A. (2018). Smart Contracts: What and Why. Retrieved from https://www.simplawyer.com/wp-content/uploads/simplawyer.com-—-Смарт-контракты.-Зачем-и-как-Demo-1.0.pdf

[9] Belikova, K. (2018). *Features of the legal regulation of the digital intellectual economy*. Retrieved from https://cyberleninka.ru/article/n/osobennosti-pravovogo-regulirovaniya-tsifrovoy-intellektualnoy-ekonomiki. DOI 10.24411/2073-3313-2018-10090

[10] Kozlovskyi, S., Baidala, V., Tkachuk, O., Kozyrska, T. (2018). Managament of the sustainable development of the agrarian sector of the regions of Ukraine. *Montenegrin Journal of Economics*, 14(4), 175-190. Retrieved from http://dx.doi.org/10.14254/1800-5845/2018.14-4.12

[11] Kozlovskyi, S., Bilenko, D., Kuzheliev, M., Lavrov, R., Kozlovskyi, V., Mazur, H., Taranych, A. (2020). The system dynamic model of the labor migrant policy in economic growth affected by COVID-19. *Global Journal of Environmental Science and Management*, 6(Special Issue (Covid-19)), pp. 95-106. Retrieved from http://dx.doi.org/10.22034/GJESM.2019.06.SI.09

[12] Bakhmat, N., Maksymchuk, B., Voloshyna, O., Kuzmenko, V., Matviichuk, T., Kovalchuk, A. … Maksymchuk, I. (2019). Designing cloud-oriented university environment in teacher training of future physical education teachers. *Journal of Physical Education and Sport*, 19(4), 1323-1332.

[13] Sitovskyi A., Maksymchuk B., Kuzmenko V., Nosko Y., Korytko Z., Bahinska O. … Maksymchuk, I. (2019). Differentiated approach to physical education of adolescents with different speed of biological development (2019). *Journal of Physical Education and Sport*, 19 (3), Art 222, pp. 1532-1543.

[14] Kozlovskyi, S., Grynyuk, R., Baidala, V., Burdiak, V., Bakun, Y. (2019). Economic Security Management of Ukraine in Conditions of European Integration. *Montenegrin Journal of Economics*, 15(3), 137-153. Retrieved from http://dx.doi.org/10.14254/1800-5845/2019.15-3.10

[15] Churchill G. A., Nevin J. R., Watson R. R. (1977). *The role of credit scoring in the loan decision*. Credit World. March.

[16] Okwonu, F. (2012). A Model Classification Technique for Linear Discriminant Analysis for Two Groups. *International Journal of Computer Science Issues*, 9, Issue 3(2), pp.125-128.

[17] Bolton, C. (2009). Logistic regressions and its application in credit scoring. University of Pretoria, 2009, 240 c.

[18] Kozlovskyi, S., Shaulska, L., Butyrskyi, A., Burkina, N., Popovskyi, Y. (2018). The marketing strategy for making optimal managerial decisions by means of smart analytics. *Innovative Marketing*, 14(4), 1-18. Retrieved from https://doi.org/10.21511/im.14(4).2018.01

[19] Kuan, C., White, H. (1994). Artificial neural networks: an econometric perspective. *Econometric Reviews*, 13, 1-91.

[20] Lavrov, R., Beschastnyi, V., Nikolenko, L., Yousuf, A., Kozlovskyi, S., Sadchykova, I. (2019). Special aspects of the banking institutions rating: a case for Ukraine, *Banks and Bank Systems*, 14(1), 48-63. Retrieved from http://dx.doi.org/10.21511/bbs.14(3).2019.05

[21] Yousuf, A., Haddad, H., Pakurar, M., Kozlovskyi, S., Mohylova, A., Shlapak, O., & Janos, F. (2019). The effect of operational flexibility on performance: a field study on small and medium-sized industrial companies in Jordan. *Montenegrin Journal of Economics*, 15(1), 47-60. Retrieved from https://ideas.repec.org/a/mje/mjejnl/v15y2019i1p47-60.html

[22] Shaulska, L., Yurchyshena, L., Popovskyi, Y. (2021). Using MS Power BI Tools in the University Management System to Deepen the Value Proposition, *11th International Conference on Advanced Computer Information Technologies (ACIT)*, Deggendorf, Germany, 294-298.

[23] Bogachev, V., Kolesnikov, A. (2012). The task of Monja-Kantorovich: achievements, connections and prospects. *Success of mathematical sciences*, 5(407), 3-110.

[24] Bucht R., Hicks R. (2018). Definition, concept and measurement of the digital economy. *Bulletin of international organizations*, 13, 143-172.

[25] European Parliament. *Challenges for Competition Policy in a Digitalised Economy*. Brussels: Euro- pean Parliament. Retrieved from http://www.europarl.europa.eu/RegData/etudes/STUD/2015/542235/IPOL_STU(2015)542235_EN.pdf/.

[26] Salchenberger L., Cinar E. & Lash N. (1992). Neural networks: a new tool for predicting thrift failures. *Decision Sciences*, 23, 899-916.

[27] Matviychuk, A., Strelchenko, I., Vashchaiev, S., & Velykoivanenko, H. (2019). Simulation of the Crisis Contagion Process Between Countries with Different Levels of Socio-Economic Development. *Proceedings of the 15th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer (ICTERI 2019)*. Retrieved from http://ceur-ws.org/Vol-2393/paper_423.pdf.

[28] Bilenko, D., Kozlovskyi, S., Ivanyuta, N., Baidala, V., Lavrov, R., Kozlovskyi, V. (2022). Efficiency assessment of tax measures in the European countries against the effects of COVID-19. *Problemy Ekorozwoju – Problems of sustainable development*, 17(1), 16-22.

[29] Tkachuk, V., Yaremova, M., Tarasovych, L., Kozlovskyi, V., Piliavoz, T. (2019). Economic Strategy of the Development of Renewable Energy in Rural Areas of Ukraine. *Montenegrin Journal of Economics*, 15(3), 71-82.

[30] Matviychuk, A., & Velykoivanenko, H. (2014). Modeling of tax incentives effectiveness. *Economic Annals-XXI*, 5-6, 105-108. Retrieved from http://soskin.info/userfiles/file/2014/5-6_2014/Matviychuk_Velykoivanenko.pdf

[31] Matviychuk, A., Novoseletskyy, O., Vashchaiev, S., Velykoivanenko, H., & Zubenko, I. (2019). Fractal analysis of the economic sustainability of industrial enterprise. *Proceedings of the Selected Papers of the 8th International Conference on Monitoring, Modeling & Management of Emergent Economy (M3E2-EEMLPEED 2019)*. Retrieved from http://ceur-ws.org/Vol-2422/paper37.pdf.

[32] Matviychuk, A. (2010). Bankruptcy prediction in transformational economy: discriminant and fuzzy logic approaches. *Fuzzy economic review*, 15(1), 21-38. DOI: 10.25102/fer.2010.01.02

[33] Derbentsev, V., Matviychuk, A., & Soloviev, V.N. (2020). Forecasting of Cryptocurrency Prices Using Machine Learning. *In L. Pichl, C. Eom, E. Scalas, & T. Kaizoji (Eds.), Advanced Studies of Financial Technologies and Cryptocurrency Markets* (pp. 211-231). Singapore: Springer. DOI: 10.1007/978-981-15-4498-9_12.

[34] Kozlovskyi, S., Fonitska, T. (2013). Modern theoretical and methodological approaches to the budget management system forming. *Economic Annals-XXI*, 3-4, pp. 35-37.

[35] Britchenko, I., Savchenko, L., Naida, I., Tregubov, O. (2020). Areas and means of formation of transport regional complexes and mechanisms for managing their competitiveness in Ukraine. *Ikonomicheski Izsledvaniathis link is disabled*, 29(3), 61-82.

[36] Matviychuk, A. (2006). Fuzzy logic approach to identification and forecasting of financial time series using Elliott wave theory. *Fuzzy economic review*, 11(2), 51-68. DOI: 10.25102/fer.2006.02.04.

[37] Matviychuk, A., Lukianenko, O., & Miroshnychenko, I. (2019). Neuro-fuzzy model of

country's investment potential assessment. *Fuzzy economic review*, 24(2), 65-88.

[38] Oliinyk, V. (2018). Optimal Management of GDP Components. *Journal of Advanced Research in Law and Economics*. IX, 2(32), 603-614. DOI:10.14505/jarle.v9.2(32).24.

[39] Babenko, V., Nehrey, M., Gaponova, E.V., Ryzhikova, N.V., & Zaporozhets, E. (2019). Life Expectancy of Population of the Country: The Role of Health Services Effectiveness. *Research in World Economy*, 10, 86-91. DOI: https://doi.org/10.5430/rwe.v10n4p86

[40] Jiang, Y., Havrysh, V., Klymchuk, O., Nitsenko, V., Balezentis, T., Streimikiene, D. (2019). Utilization of Crop Residue for Power Generation: The Case of Ukraine. *Sustainability*, *11*, 7004. https://doi.org/10.3390/su11247004

[41] Voynarenko M, Dzhedzhula V., Yepifanova I. (2020). Modeling of the process of personnel motivation for innovation activity. *WSEAS Transactions on Business and Economics*, 17, 467-477.

[42] Kaletnik G., Pryshliak N., Khvesyk M., Khvesyk J. (2022). Legal regulations of biofuel production in Ukraine. *Polityka Energetyczna*, 25(1), 125-142. DOI: https://doi.org/10.33223/epj/146411. https://epj.min-pan.krakow.pl/Legal-regulations-of-biofuel-production-in-Ukraine,146411,0,2.html

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflict of interest to declare that is relevant to the content of this article.