УДК 338.27

Romanovskyi I.
Candidate of Science (tech.), Associate Professor,
Associate Professor of Economy and Entrepreneurship Department of
National Metallurgical Academy of Ukraine

APPLICATION OF MOST-LIKELIHOOD METHOD FOR SALES FORECAST AND ITS TESTING IN THE TERMS OF AN INDUSTRIAL ENTERPRISE

ЗАСТОСУВАННЯ МЕТОДУ НАЙБІЛЬШОЇ ВІРОГІДНОСТІ ДЛЯ ПРОГНОЗУВАННЯ ПРОДАЖІВ ПРОМИСЛОВОГО ПІДПРИЄМСТВА

ANNOTATION

In conditions of stiff market competition, strict requirements are being set to forecasting methods to secure reliable results of actual sales quantity prediction. The characteristic feature of sales forecasting methods is the use of time series. While taking into account the influence of time factor on sales (in straight-line, exponential forms, etc.), this approach does not account for and miss a number of factors that affect the sales. Moreover, a seller of goods must focus on factors that influence the decision of a buyer to buy his product (not his competitors'). Taking into account such features not only improves strategic planning, but also makes it possible to gain additional benefit due to impact on the buyer as the result of upgrading his sales process. An original model based on most likelihood method was created to solve the issue of sale forecast. The application of such a model has allowed to quantify an impact of each individual factor on probability of sales based on historical data. The most-likelihood method was applied to find multipliers of the logistic regression equation. Testing of the model in the terms of an industrial enterprise proved to be efficient. The sensitivity and specificity tests of the model affirmed its reliability. Basing on the results of calculations to estimate reliability, the ROC-curves were plotted for the basic and optimal values of logistic regression coefficients. Due to the model additional advantages over the competitors of the enterprise will be received.

Keywords: sales forecast, model, most-likelihood method, coefficients, reliability, advantages, enterprise.

АНОТАЦІЯ

Визначені вимоги до методів прогнозування щодо забезпечення надійних результатів визначення фактичної кількості продажів. З'ясовано, що характерною рисою більшої методів прогнозування продажів є використання часових рядів. Враховуючи вплив фактору часу на продажі (у прямолінійних, експоненційних формах тощо), цей підхід не розглядає та пропускає низку суттєвих факторів, які впливають на обсяги продажів. Існує потреба в інструментарії, завдяки якому продавець товарів може зосередитись на факторах, що впливають на рішення покупця купувати саме його товар. Врахування таких особливостей покращує стратегічне планування, дає можливість отримати додаткову вигоду через вплив на покупця в результаті вдосконалення товару та процесу продажу безпосередньо. З метою прогнозування продажів промислового підприємства пропонується модель, застосування якої дозволяє на основі історичних даних кількісно оцінити вплив кожного окремого фактору на вірогідність продажу. Для визначення множників рівняння логістичної регресії було застосовано метод найбільшої вірогідності. Тестування моделі в умовах промислового підприємства виявилося ефективним. Результати випробувань чутливості та специфічності моделі підтвердили його надійність. На підставі результатів розрахунків для оцінки надійності були побудовані криві помилок для базових та оптимальних значень коефіцієнтів логістичної регресії. Модель дозволяє отримувати додаткові переваги над конкурентами підприємства.

Ключові слова: прогноз продаж, продаж, модель, можливості методу, коефіцієнти, надійність, переваги, підприємство.

АННОТАЦИЯ

Определены требования к методам прогнозирования по обеспечению надежных результатов определения фактического количества продаж. Установлено, что характерной чертой большей методов прогнозирования продаж является использование временных рядов. Учитывая влияние фактора времени на продаже (в прямолинейных, экспоненциальных формах и т.п.), этот подход не рассматривает и упускает ряд существенных факторов, влияющих на объемы продаж. Существует потребность в инструментарии, благодаря которому продавец товаров может сосредоточиться на факторах, влияющих на решение покупателя покупать именно его товар. Учет таких особенностей улучшает стратегическое планирование, дает возможность получить дополнительную выгоду через воздействие на покупателя в результате совершенствования товара и повышении эффективности процесса продажи. С целью прогнозирования продаж промышленного предприятия предложена математическая модель, основанная на методе наибольшего правдоподобия. Использование такой модели позволяет на основе исторических данных количественно оценить влияние каждого отдельного фактора на вероятность продажи. Тестирование модели в условиях промышленного предприятия оказалось эффективным. Результаты испытаний чувствительности и специфичности модели подтвердили ее надежность. На основании результатов расчетов для оценки надежности были построены кривые ошибок для базовых и оптимальных значений коэффициентов логистической регрессии. Модель позволяет получать дополнительные преимущества над конкурентами предприятия.

Ключевые слова: прогноз продаж, продажа, модель, возможности, коэффициент, надежность, преимущества, предприятие.

Formulation of the problem. Sales forecasting is an important component of the management system of an industrial enterprise [1]. Effective and thorough forecasting has a significant impact not only on the strategic plans of enterprise development, but on the result of its work in general. Therefore, in conditions of stiff market competition, additional requirements are being set to forecasting methods to secure reliable results of actual sales quantity prediction. Such methods should take into account both the features of enterprise and the market it works in, as well as to envisage a possibility of operational adjustments and revisions, taking into account the slightest changes in the market environment. Therefore, research on the sales forecasting issue with application of mathematical methods and modern software is of paramount importance. It tends to be relevant, demand-driven and promising.

Analysis of recent research and publications. A large number of researchers investigate the sales forecasting issues. Depending on the industry and the purpose of application, a wide range of methods have been used to solve this problem. The main methods of forecasting include heuristic and mathematical (quantitative) ones [2].

Heuristic methods are usually based on a jury of expert (manager) opinion. These methods include method of generating ideas; simple ranking method, weighting method, sequential comparison method, paired comparison method etc. A special attention has been drawn to Delphi method, which is based on a jury of expert opinion as well [3].

Among the mathematical methods, there are three groups [2]:

– simple extrapolation methods in time series, including – method of least squares, exponential smoothing, etc.;

– statistical methods – correlation and regression analyzes, factor analysis, etc.;

– combined methods – synthesis of different variants of forecasts.

The statistical methods of forecasting sales are of the most and widespread practical use. These methods investigate cases when each value of a variable (actual sales quantity) is matched by a certain distribution of another variable (time series). Such correspondence is usually referred to as stochastic dependence.

Application of statistical dependence to forecasting sales is due to the fact that the dependent variable falls under the influence of uncontrollable factors.

Due to uncertainty of statistical dependence between the two values (an undependable and dependable ones), a mean value of dependable parameter is widely used in the form of equation of mathematical expectation.

In general, in forecasting sales, correspondence of each value of the time series to certain arbitrary mathematical expectation of actual sales quantity is widely used. It can be set as a correlation function

$$M_x(Y) = F(X_i), \quad (1)$$

Where Y – a feedback function;

X_i –an undependable value.

The application of regression analysis to tackle the forecasting sales problem is aimed at determining the dependence of random variables (time series) on non-random variables (actual sales quantity).

Given the relationship between parameters can be assumed to have a linear regression model

$$Y = \beta_o + \beta_1 + \varepsilon, \quad (2)$$

the regression equation is to be sought in the form of a linear equation

$$y = b_0 + b_1 x \quad (3)$$

According to the least squares method (LSM), the unknown parameters b_0 and b_1 are chosen under conditions that the sum of squares of deviations of empirical values y_i from the theoretical values of $\hat{y}_i$ is the smallest:

$$S = \sum_{i=1}^{n} (\bar{y} + y_i)^2 = \sum_{i=1}^{n} (b_0 + b_1 x_1 - y_i)^2 \to min \quad (4)$$

Correlation coefficient is a measure of the strength and direction of the linear relationship between two variables

$$r = \frac{\overline{xy} - \bar{x}\,\bar{y}}{S_x S_y} \quad (5)$$

where S_x, S_y – standard deviations.

In the case where the time sales function is not linear, methods based on other dependencies of time parameter change are used (for example, expositional smoothing).

A certain advantage of correlation analysis is ability to take into account the influence of several factors (not only time). In this case, multi-factor correlation analysis should be involved.

As for application of regular correlation analysis to economic problems, it should be applied only when the factors being analyzed are given in usual numerical form. If a phenomenon, process, or factor under examination takes other forms (for example, binary values), then application of ordinary regression apparatus would be erroneous. To solve the problem another statistical tool must be applied.

Highlighting the previously unsettled issues the article is devoted to.

The characteristic feature of sales forecasting methods is the use of time series. While taking into account the influence of time factor on sales (in straight-line, exponential forms, etc.), this approach does not account for and miss a number of factors that affect the sales. Moreover, a seller of goods must focus on factors that influence the decision of a buyer to buy his product (not his competitors'). Taking into account such features not only improves strategic planning, but also makes it possible to gain additional benefit due to impact on the buyer as the result of upgrading his sales process.

Formulating the goals of the article (task statement)

In the opinion of the author, the most appropriate and objective approach to this issue is creating of a probabilistic model of forecasting sales. The application of such a model allows to quantify an impact of each individual factor on probability of sales based on historical data.

The article is aimed at creation, substantiation and testing of a model of sales forecasting taking into account the requirements of consumers, specificity of products etc. The use of such a model gives a manufacturer (seller) benefits because it enables him to identify the priority directions of product improvement in order to strengthen the market positions. In addition, the use of such a model should take into account the impact on sales not only of properties of the product, but also services that accompany the sale and after-sale service.

The model for forecasting sales, along with taking into account a wide variety of factors, should be easily used in practical terms.

Presentation of the main research material with full substantiation of the received scientific results.

To determine the factors that have the most significant effect on the probability of purchasing products from its manufacturer – the industrial enterprise – a classic statistical toollkit was considered, by which one can estimate the degree of influence of each individual factor [4; 5] on the decision-making of consumers. Since the result of the purchase of products from a manufacturer has a binary nature (the product was bought / wasn't bought), a toolkit linking quantitative and attributive parameters for construction of the mathematical model was chosen.

Such a toolkit is logistic regression [4], which is a kind of multiple regression, the general purpose of which is to analyze the relationship between dependent variable and several independent variables (predictors). Since the binary logistic regression is used in a case when a dependent variable accepts one of the two possible values (yes or no), the situation turns into a binary decision – a consumer either buys the product or fails to buy it.

Based on the study of the sales department of an industrial enterprise (not all applicants in the end are buying goods), an event of the subsequent purchase is considered to be marked «1», the refusal – as «0». Later, the relative frequency of occurrence of each events in this binary system has been used to calculate the probability of occurrence of the event «1» (purchase of products).

As is known, all regressive models can be written as [6]:

$$y = F(x_1, x_2, x_3 \dots x_n) \qquad (6)$$

For multiple linear regression, it is assumed that the dependent variable is a linear function of independent variables [6]:

$$y = b_0 + b_1x_1 + b_2x_2 + \dots b_nx_n \qquad (7)$$

When we try to use usual multiple regression to solve the problem, the following problem arises: multiple regression does not take into account the response of binary variables, which leads to probabilities greater than 1 and less than 0, that is erroneous [7].

To solve the problem, the regression problem was formulated as follows: instead of a binary variable, a continuous variable for the interval [0,1] was involved. To achieve our goal, we applied a regression equation in a form of logit function [8]

$$P(E) = \frac{1}{1 + EXP\left(-\left(b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n\right)\right)} \qquad (8)$$

where $P(E)$ – the probability that the event happens;

e – the basis of natural logarithms

$b_0, b_1, b_2 \dots b_n$ – standard multipliers of regression equation,

$x_0, x_1, x_2 \dots x_n$ – values of 1, 2, 3 ... n-th factors of the regression equation.

We apply most-likelihood method [9; 10] (hereinafter referred to as MMP) to find multipliers of the logistic regression equation. It is based on the function of probability (likelihood function or LL-function) [8], which expresses density of the probability function (PDF) of the conjugate appearance of results of the sample $Y_1, Y_2, \dots, Y_k$:

$$logit(P) = log_e \frac{P(E)}{1 - P(E)} \qquad (9)$$

$$L(Y_1, Y_2 \dots Y_n; \theta) = P(Y_1; \theta) \dots P(Y_n; \theta) \qquad (10)$$

In accordance with MLM, the value of the function $\theta = \theta(Y_1, Y_2 \dots Y_n)$ is taken as an estimator of the unknown parameter, which maximizes the LL-function. To simplify the decision process, we need to maximize the natural logarithm of the LL-function, since the maximum of both functions is achieved with the same value of θ [8]:

$$L(Y; \theta) = \ln\left(L(Y; \theta)\right) \to max \qquad (11)$$

Thus, the logarithmic probability function is equal to

$$LL = lnL = \sum_{i=1}^{n} [y_i \ln p_i + (1 - y_i) \ln(1 - p_i) \qquad (12)$$

Where p_i – a set of observations for which $y_i = 1$

Values of function (7) are always negative due to its permanent concave form [9].

Gradient methods are usually applied to determine the coefficients of logistic regression.

To assess the quality of a model based on logistic regression, it is convenient to use – ROC analysis, that is, to create an error curve [9].

ROC-curve (Receiver Operator Characteristic) belongs to a family of curves of signal processing systems, which are most often used to represent the results of binary classification in computer simulation [9; 10]. One of the two possible classes (0 or 1) is defined as a class with positive effects, the other one (0) – with negative ones. Involving the ROC curve proves to be efficient in distinguishing of correct classified negative examples from incorrectly classified ones (respectively, a truly positive and false negative set).

Using a special parameter (called discrimination threshold or cut-off value) varying from 0 to 1, we obtain an appropriate division of the set into two classes, which depends on the different magnitudes of errors of type I and II.

To demonstrate the nature of the errors of type I and II, a confusion matrix has been complied. Table 1 represents the data based on the results of model classification and the actual attachment of examples to classes [11].

In Table 1 the following definitions are used. TP- correctly classified positive examples (true positive cases); TN – correctly classified negative examples (true negative cases); FN – posi-

Table 1

Conjugation table

Model (classified as)	True condition	
	positive	negative
positive	TP	FP
negative	FN	TN

tive examples, classified as negative (Error type I), FP – negative examples, classified as positive (Error type II) [11].

Classification of events as positive and negative is conditional and depends on the aim of particular task. In our case, the fact of purchasing products of the enterprise by each individual consumer is classified as positive. A refusal to purchase in case of previous interest in the goods, but the negative one is (commercial request or inquiry for delivery of products in the sales department of the enterprise).

In the data analysis process we operate relative indicators – rates. The rate of truly positive examples (True Positives Rate)

$$TRP = \frac{TP}{TP + FN} \times 100\% \quad (13)$$

False Positives Rate equals

$$FPR = \frac{FP}{TN + FP} \times 100\% \quad (14)$$

To determine the objective value of our binary model, we additionally use indicators of sensitivity and specificity of the model.

Sensitivity is defined as a rate of truly positive cases:

$$SE = TRP = \frac{TP}{TP + FN} \times 100\% \quad (15),$$

and Specificity – as a rate of truly negative cases that were correctly identified by the model:

$$SP = \frac{TN}{TN + FP} \times 100\% \quad (16)$$

It should be noted that

$$FPR = 100\% - SP \quad (17)$$

For approbation and evaluation of reliability of the mathematical model for determining the impact of individual parameters on the sales of industrial enterprise (the number of sales operations), sales database on an industrial enterprise was examined. Information on approximately 4500 buyers' inquiries was processed, out of which 59% inquiries ended up buying the products, 41% inquiries were cancelled.

The factors listed in Table 2 were used to assess quality of the model. The The same identifies which factors are quantified, and which factors have attribute values; the multipliers of logit function (5) matching the most-likelihood function (12) maximum value found with help of our model are shown in Table 2 as well.

Calculations of regression coefficients were performed using the EXCEL software, the «Solver» function, which performs iterative calculations using the generalized reduced gradient method (GRG) [5,9].

Basing on the results of calculations to estimate reliability, the graphs of error curves (ROC-curve) were plotted for the basic and optimal values of logistic regression coefficients. The results are shown in Figures 1 and 2.

The analysis of the results of calculations of the functions of maximum probability (LL) shows that in optimum terms, compared to the basic, the accuracy of the model increases, since the value of LL increases from (-321.55) in the base version to (-202.24) with new regression coefficients.

In order to conclude whether the optimal plan is more accurate compared to the baseline, it is necessary to compare the error curves (ROCs) for the baseline and optimal plans. A visual comparison of the ROC curves barely allows us to identify the most efficient model. A reliable method for comparing ROC curves is to estimate the area under

Table 2

Sales database parameters

№	Parameter	Value's feature	Logit function multipliers
1	Enterprise's product price to average market price ratio	continuous variable	0.4537
2	Number of items in the order	continuous variable	0.0330
3	Term of the order's fulfilment	continuous variable	0.1432
4	Term of the order processing	continuous variable	0.0275
5	After-sale service	attribute value	-0.3608
6	Delivery of goods to the consumer	attribute value	0.0910
7	Availability of additional properties of the product (in comparison with technical documentation requirement)	attribute value	0.2976
8	Vendor-buyer cooperation period duration	continuous variable	-0.2363
9	Buyer Distribution Zone	attribute value	-0.1934
10	Availability of personal discounts	attribute value	0.1252
11	Work with a personal manager	attribute value	0.3772
12	Form of ownership of the enterprise-customer	attribute value	0.1571
13	Features of the order (at the head office or through a sales representative)	attribute value	-0.0589
14	Form of cash settlements	attribute value	0.1752

the curves. In practice, it usually varies between 0.5 («false classifier») and 1.0 («ideal model»).

Comparison of calculated values for optimal AUC_{OPT} = 0.6764 and baseline AUC_{BASE} = 0.54162 reaffirms the greater reliability of the optimal plan, that is, the calculated values of probability, which are determined with the help of the coefficients of logistic regression we have obtained.

The ideal model has 100% sensitivity and specificity, but it is unachievable to keep sensitivity and specificity of the model simultaneously at the highest level. To solve the problem, the discrimination threshold (cut-off point) is used.

The issue of finding a optimal cut-off point could also be set forth. In our case, 0.75 cut-off point was selected (75% of probability that the product will be bought).

The AUC value for the optimal plan is 0. 6764, testifies that more than two thirds of cases are determined correctly with the probability of 75% when determined using the proposed logistic regression equation.

Based on the values of the coefficients of regression from Table 2, we formulate the sales forecasting probability function putting the logit function multipliers (Table 2) into the equation (7). Finally, putting the client featuring parameters $(x_1, x_2 \ldots x_n,)$ – into equation (7) we obtain the precise value of probability function determining (at the level of cut-off point) whether the client buys the product or fails to do it.

Moreover, due to equation (7) and the logit function multipliers (Table 2), the producer (seller) can quantify the effect of each factor included into the probability function on actual sales quantity. As far as to find out what should be done to enhance attractiveness of its product. It provides additional market advantages over competitors.

Conclusions from this study and prospects for further exploration in this direction

On the base of the analysis of works examining the forecasting of sales, it was found that the main attention is drawn to forecasting sales using time series.

To a significant degree the issue of studying the characteristics of consumers that affect the sale of products of an industrial enterprise is put aside.

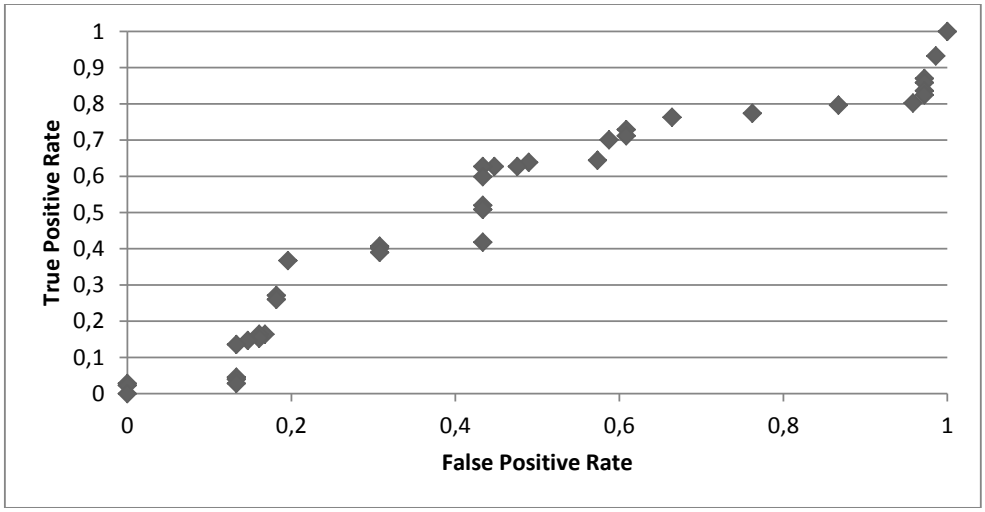

Fig. 1. The error curve for the base version

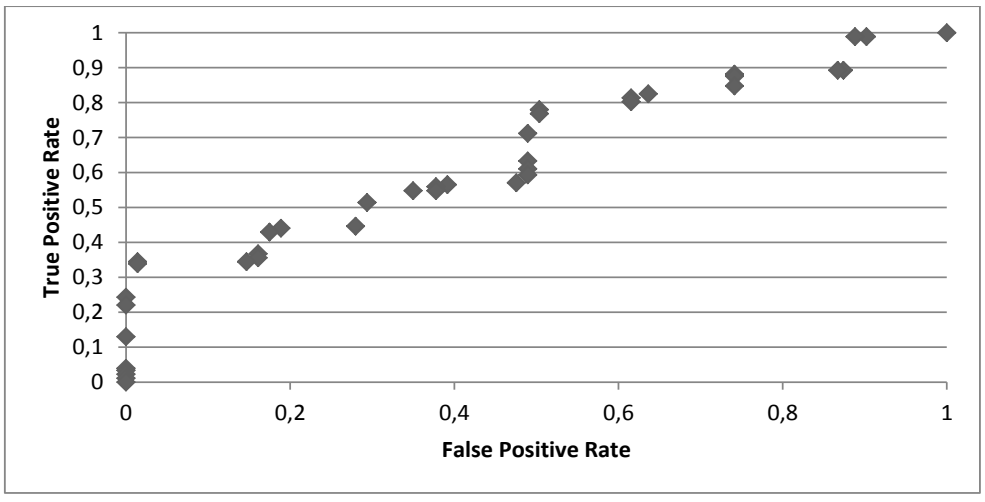

Fig. 2. The error curve for the optimal plan

For the quantitative estimation of the projected volume of sales, mathematical apparatus is involved, most frequently – the correlation analysis.

The author proposes a mathematical model that allows to forecast sales using the probability of sales function. To create the equation of this function, logistic regression is implied. It enables to take into account the influence of attributive and quantitative parameters simultaneously.

The model set forth by the author allows evaluating the impact of each individual factor on the sales probability. Doing so would permit to identify priority areas for improving the product and services offered by the seller (producer) in order to enhance the attractiveness of its product. Testing of research results in the terms of a industrial enterprise and proved to be reliable.

As possible directions of the research, the author considers to involve cluster analysis to process initial data as well as envisages development of a mathematical model to solve problems that, unlike binary response (yes / no), have multivariable results.

REFERENCES:

1. Підприємницька діяльність та економіка підприємства: навч. посіб. / С.Б. Довбня, Т.Б. Ігнашкіна, А.Б. Педько та ін.; за заг. ред. д-ра екон. наук, проф. С.Б. Довбні. – Д.: ЛІРА, 2016. – 440 с.
2. Castells M.A. DIRRECCION DE VENTAS – Organizacion del departamento de ventas y gestion de vendedores / Castells. Manuel Artal Castells. – Madrid. : ESIC EDITORIAL. 2016 – 554 p.
3. Chiesa C. Los pecados capitales de la venta: 40 errores a evitar en su estrategia commercial / Cosimo Chiesa. – Barcelona: Empresa Activa. 2010. – 138 p.
4. OpenIntro Statistics Second Edition [Електронний ресурс]: web-сайт. – Режим доступу https://www.openintro.org/stat/labs.php/
5. Matrices de Correlaciones y Similaridades/disimilaridades [Електронний ресурс]: web-сайт. – Режим доступу https://www.xlstat.com/es/ soluciones/funciones/matrices-de-correlaciones-y-similaridades-disimilaridades
6. Бідюк П. І. Прикладна статистика / П. І. Бідюк, О. М. Терентьєв, Т. І. Просянкіна-Жарова. – Вінниця : ПП ТД «Едельвейс і К», 2013. – 304 с.
7. Бобик, О. І. Теорія ймовірностей і математична статистика / О. І. Бобик, Г. І. Берегова, Б. І. Копитко. – К. : Професіонал, 2007. – 560 с.
8. In Jae Myung. Tutorial on maximum likelihood estimation/ Journal of Mathematical Psychology, Volume 47, Issue 1, February 2003, Pages 90–100.
9. Franses, Ph. H. Time series models for business and economic forecasting / Ph. H. Franses. – Cambridge : Cambridge University Press, 1998. – 280 p.
10. Эфрон, Б. Нетрадиционные методы многомерного статистического анализа / Б. Эфрон. – М. : Финансы и статистика, 1988. – 263 с.
11. Curva ROC Wikipedia [Електронний ресурс]: web-сайт. – Режим доступу: https://es.wikipedia.org/wiki/Curva_ROC