УДК 681.518.5

Prof. **Khlaponin Y.I. (KNUCA)**
Ph.D. **Qasim N. H. (KNUCA)**
Tarasiuk D.M. (KNUCA)

DOI: http://doi.org/10.17721/2519-481X/2023/80-10

FORECASTING THE STATE OF TELECOMMUNICATION NETWORKS USING QUANTILE AND LOGISTIC REGRESSION METHODS

In today's modern world, the ubiquity of information technology has intertwined telecommunications systems with every facet of human life. It's challenging to fathom a world where you're disconnected from the "World Wide Web" or unable to exchange data instantly via the intricate web of modern mobile devices. The vitality of staying connected online cannot be overstated, and ensuring the smooth functioning of telecommunications systems is paramount. This paper delves into the pivotal task of predicting and managing the performance of these networks, employing quantile and logical regression techniques. Our study leverages real-world data from telecommunication network operations to construct a predictive model capable of anticipating network conditions in advance. This predictive capability serves as the linchpin for intelligent decision-making systems, facilitating real-time network management. By implementing machine learning methods, specifically quantile regression, we achieve a sophisticated understanding of how various factors influence network performance. Our research doesn't just stop at forecasting; it extends to the realms of intellectual decision-making systems, where the insights gained from regression analysis play a pivotal role. These intelligent systems are equipped to make data-driven decisions on network resource allocation, maintenance schedules, and preemptive problem resolution. In essence, they act as the custodians of network stability, ensuring that telecommunication systems remain robust and responsive to the ever-evolving demands of modern society. This article sheds light on the indispensable role of regression methods in proactively managing the state of telecommunications networks. By harnessing the power of machine learning and data-driven insights, we pave the way for a future where network disruptions are minimized, and the seamless connectivity we've come to rely on remains a constant presence in our lives.

Keywords: information technology, telecommunication system, intellectual decision-making system, machine learning, quantile regression.

Introduction. Modern life is hard to imagine without information technology. A set of numerous digital radio-electronic devices connected to telecommunication systems not only help in organizing business processes at large enterprises, but also deeply entered our daily life. It is difficult to overestimate the importance of being in touch 24 hours a day and exchanging data at any second through networks consisting of modern mobile devices. Despite the fact that at first glance, many electronic devices seem to be independent, upon closer examination, they perform the functions of collecting, processing and transmitting information. And even if the device provides all of the above functionality, they still need to exchange information with each other through telecommunication networks.

At the moment, information technology has gone far ahead, which makes it possible to transfer information between devices at a high level. However, due to the complex multifunctional nature of these devices, the issues of maintaining high efficiency of information exchange play a key role in information technology. One of the main characteristics that determine the quality of the functioning of digital radio-electronic devices, designed solely to support the exchange of data between the remaining devices, is reliability. The operability and reliability of telecommunication systems and computer networks, consisting of radio-electronic devices for processing and transmitting information, will ultimately be characterized by the operability and reliability of the most inefficient device. Therefore, timely warning by identifying relevant inefficient devices is a key element in the analysis and monitoring of such systems.

Recently, to solve the problems of analysis and monitoring of telecommunication systems and computer networks, applied intelligent technologies are increasingly used [7]. Such technologies are based on the so-called machine learning technologies, on the basis of which decision-making systems are built for the intelligent control of complex distributed info communication

networks [1]. Machine learning algorithms designed for predictive use basically use the idea of autocorrelation between values over time. The general structure of the intelligent control system:

• telecommunications network;
• database of characteristics of the telecommunications system;
• data processing system based on machine learning technologies;
• decision making system;

Intelligent systems can differ significantly in their functions, but they always contain these blocks to one degree or another [2].

It is important to note that the main architectural feature that distinguishes the intelligent control system (Fig. 1.) from the one built according to the "traditional" scheme is due to the connection of storage mechanisms and intelligent data processing that characterize the operation of systems.

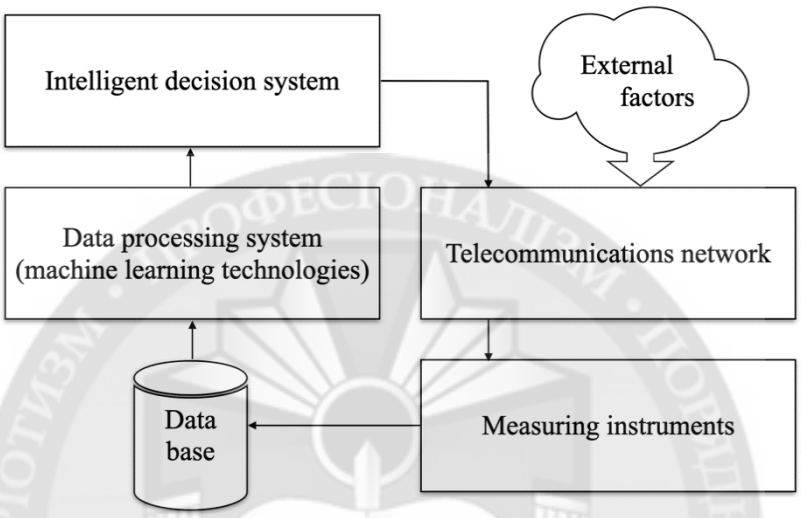

Figure - 1. General structure of the intelligent control system

This feature of intelligent systems allows real-time monitoring, analysis and effective management of the telecommunications network.

Formulation of the problem. The main purpose of the study is to build a mathematical model for predicting indicators that characterize the operation of telecommunication networks in order to optimize an intelligent decision-making system.

Research method. In this work, the quantile regression method was used as a machine learning model. In this work, the theory of logistic regression was used as a machine learning model.

Logistic regression [3, 4] is used to predict the probability of occurrence of some event based on the values of a set of features. To do this, the so-called dependent variable is introduced, which takes only one of two values - as a rule, these are the numbers 0 (the event did not occur) and 1 (the event occurred), and a set of independent variables (also called features, predictors or regressors) - real $x_1, x_2, ..., x_n$, based on the values of which it is required to calculate the probability of accepting one or another value of the dependent variable.

It is assumed that the probability of an event occurring $y = 1$ equal:

$$\mathrm{P}\{y = 1 \mid x\} = f(z),$$

where $z = \theta^T x = \theta_1 x_1 + ... + \theta_n x_n$, x and θ column vectors of values of independent variables $x_1, x_2, ..., x_n$ and parameters (regression coefficients) - real numbers $\theta_1, \theta_2, ..., \theta_n$, respectively, $f(z)$– the so-called logistic function (sometimes also called the sigmoid or logit function):

$$f(z) = \frac{1}{1 + e^{-z}}.$$

Since it takes only the values 0 and 1, the probability of the first possible value is equal to:

$$\mathrm{P}\{y = 0 \mid x\} = 1 - f(z) = 1 - f(\theta^T x).$$

For brevity, the distribution function for given can be written in the following form:

$$\mathrm{P}\{y \mid x\} = f\left(\theta^T x\right)^y \left(1 - f\left(\theta^T x\right)\right)^{1-y}, \quad y \in \{0,1\}.$$

Advantages of the quantile regression method. The need to develop new methods of statistical estimation is dictated by the practical need for more adequate mathematical tools. Among the applied regression methods, the most common is the least squares method, which allows obtaining deep statistical results under the assumption that random errors are distributed according to a normal (Gaussian) law and (in most cases) are independent. There are also alternative approaches that require, however, strict assumptions about the type of distributions and other requirements regarding observations.

However, in practice, one often has to deal with more complex situations that do not fit into the standard assumptions of regression methods. Examples of such "violations" include:

1. the inaccuracy of setting the distribution that controls the observations in the sample, so that the assumptions of classical regression models about the homogeneity of the sample or about a sufficiently “beautiful” mechanism for the manifestation of homogeneity cannot be verified;

2. the presence of distributions with more "heavy tails" than the normal distribution, which necessitates the choice of estimation methods that give less weight to the extreme observed values, or even a complete rejection of the least squares method;

3. the presence in the sample of a small proportion of "outliers", that is, observations caused by some kind of "noise", which, as a rule, cannot be separated on the basis of a priori information. This requires the use of procedures that are not very sensitive to such “contamination” of the sample;

4. the dependence of the elements of the sample, which has a complex structure, so that it is difficult or even impossible to isolate and / or analyze (for example, using a covariance matrix).

In other words, in a number of practical problems that require the use of a regression approach, classical regression methods are inoperable and do not allow drawing correct conclusions about the nature of the process under study or the behavior of the object under consideration.

There have been repeated attempts to build alternative approaches. In particular, the so-called robust methods, which are resistant to deviations from the assumptions of the classical theory [5, 6], have been actively developed.

One of these approaches is the quantile regression method [8, 9, 10], which consists in replacing square deviations with absolute ones. It has a number of advantages:

–resistant to "outliers", which are often encountered in practical tasks;

–does not require independence or weak dependency;

–allows you to directly draw conclusions about the fluctuations of the predicted (estimated) indicator.

Thus, this method allows us to overcome the shortcomings of classical regression models, which are very sensitive to violations of their assumptions.

Selection of parameters. For selection of parameters $\theta_1, \theta_2, ..., \theta_n$ it is necessary to make a training sample consisting of sets of values of independent variables and the corresponding values of the dependent variable y. Formally, this is the set of pairs $\left(x^{(1)}, y^{(1)}\right), ..., \left(x^{(m)}, y^{(m)}\right)$, where $x^{(i)} \in \Re^n$ – vector of values of independent variables, and $y^{(i)} \in \{0,1\}$ – their corresponding value y. Each such pair is called a training example.

The maximum likelihood method is usually used, according to which the parameters are chosen θ, maximizing the value of the likelihood function on the training sample:

$$\hat{\theta} = \arg\max_\theta L(\theta) = \arg\max_\theta \prod_{i=1}^{m} \mathrm{P}\left\{y = y^{(i)} \mid x = x^{(i)}\right\}.$$

Maximizing the likelihood function is equivalent to maximizing its logarithm:

$$\ln L(\theta) = \sum_{i=1}^{m} \log \mathrm{P}\left\{y = y^{(i)} \mid x = x^{(i)}\right\} = \sum_{i=1}^{m} y^{(i)} \ln f\left(\theta^T x^{(i)}\right) + (1 - y^{(i)}) \ln\left(1 - f\left(\theta^T x^{(i)}\right)\right).$$

To maximize this function, for example, the gradient descent method can be applied. It consists in performing the following iterations, starting from some initial parameter value θ:

$$\theta := \theta + \alpha \nabla \ln L(\theta) = \theta + \alpha \sum_{i=1}^{m} \left(y^{(i)} - f\left(\theta^T x^{(i)}\right) \right) x^{(i)}, \alpha > 0.$$

General quantile regression model. Let $(y_i, \mathbf{x}_i)$ – set of observations ($i = 1 \div n$), where y_i – dependent variable in regression equation, and $\mathbf{x}_i = (x_{i1} \ldots x_{im})$ – row vector of independent variables (covariate). Then the model is given by the relation [8, 9]

$$Quant_\theta(y_i \mid \mathbf{x}_i) = \mathbf{x}_i \boldsymbol{\beta}_\theta, \quad (1)$$

where $Quant_\theta(y_i \mid \mathbf{x}_i)$ denotes a conditional quantile y_i for probability θ on the regressor vector $\mathbf{x}_i$, and $\boldsymbol{\beta}_\theta$ – corresponding column vector of regression coefficients.

In other words, if *i* observation is described by a random vector $(\tilde{y}_i, \tilde{\mathbf{x}}_i)$, then the solution of an optimization problem of the form

$$\inf_{\boldsymbol{\beta}_\theta} \left\{ \mathbf{x}_i \boldsymbol{\beta}_\theta \;\middle|\; F_{\tilde{y}_i \mid \tilde{\mathbf{x}}_i}\left(\mathbf{x}_i \boldsymbol{\beta}_\theta \mid \tilde{\mathbf{x}}_i = \mathbf{x}_i\right) \geq \theta \right\}.$$

Such a "direct" method requires knowledge of the conditional (joint) distribution or suitable assumptions about it.

Therefore, a nonparametric approach based on a large number is often used *n* notice $(y_i, \mathbf{x}_i)$, $i = 1 \div n$. Within its framework, the assessment $\hat{\boldsymbol{\beta}}_\theta$ vector $\boldsymbol{\beta}_\theta$ from relation (1) is obtained by solving the minimization problem:

$$\min_{\boldsymbol{\beta}_\theta} \frac{1}{n} \left\{ \sum_{i: y_i \geq \mathbf{x}_i \boldsymbol{\beta}_\theta} \theta \left| y_i - \mathbf{x}_i \boldsymbol{\beta}_\theta \right| + \sum_{i: y_i < \mathbf{x}_i \boldsymbol{\beta}_\theta} (1-\theta) \left| y_i - \mathbf{x}_i \boldsymbol{\beta}_\theta \right| \right\} \quad (2)$$

For $\theta = \frac{1}{2}$ it is reduced to its special case - the classical problem of the least distances (LAD) [11, 12]

$$\min_{\boldsymbol{\beta}_\theta} \frac{1}{n} \sum_i \frac{1}{2} \left| y_i - \mathbf{x}_i \boldsymbol{\beta}_\theta \right|$$

Representation of quantile regression as a linear programming problem. Problem (2) can be reduced to a linear programming problem of the form:

$$\begin{cases} \theta \cdot \mathbf{1} \cdot \mathbf{u}^+ + (1-\theta) \cdot \mathbf{1} \cdot \mathbf{u}^- \to \min \\ \mathbf{X}\boldsymbol{\beta}_\theta + \mathbf{u}^+ - \mathbf{u}^- = \mathbf{y}, \\ \mathbf{u}^+ \geq \mathbf{0}, \\ \mathbf{u}^- \geq \mathbf{0}, \end{cases} \quad (3)$$

where $\mathbf{1}$ – row vector of suitable dimension, consisting of ones; $\mathbf{X}$ – covariate observation matrix (dimensions $n \times m$); $\mathbf{y}$ – vector of independent variable observations (dimensions *n*); $\mathbf{u}^+$ and $\mathbf{u}^-$ – vectors of positive and negative deviations, respectively, with components

$$u_i^+ = \left(y_i - \mathbf{x}_i \boldsymbol{\beta}_\theta\right)^+ = \begin{cases} y_i - \mathbf{x}_i \boldsymbol{\beta}_\theta, \; y_i \geq \mathbf{x}_i \boldsymbol{\beta}_\theta, \\ 0, \text{ в противном случае,} \end{cases}$$

$$u_i^- = \left(\mathbf{x}_i \boldsymbol{\beta}_\theta - y_i\right)^+ = \begin{cases} \mathbf{x}_i \boldsymbol{\beta}_\theta - y_i, \; y_i < \mathbf{x}_i \boldsymbol{\beta}_\theta, \\ 0, \text{ в противном случае.} \end{cases}$$

In practice, as the initial value of the vector $\hat{\boldsymbol{\beta}}_\theta$ to solve a linear programming problem, it is convenient to take the corrected least squares estimate. Another approach is that the initial value can be obtained from a quantile regression based on a small subset of the sample, which, as a result, significantly reduces the number of iterations and computation time.

Presenting a quantile regression model as a linear programming problem has several important implications. First, it is guaranteed that the estimate will be obtained in a finite number of iterations. Second, the parameter vector estimate will be robust to outliers. In other words, if $y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}_\theta > 0$, then y_i can be increased to almost $+\infty$, and vice versa if $y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}_\theta < 0$, then y_i can be reduced to almost $-\infty$ without changing the decision $\hat{\boldsymbol{\beta}}_\theta$.

Confidence intervals for quantile regression estimates. General instructions for constructing confidence intervals for quantiles can be found in the handbook [13]. Building confidence intervals (confidence bands) for quantile regression is their development. In particular, we can name the direct method, which does not depend on the distribution, the method of studentized intervals, and the method based on the bootstrap distribution [14].

The article uses a direct approach, since it is the most optimal, both in terms of minimizing computational procedures and the adequacy of the interval estimate in general. According to this method, the confidence band is estimated for an arbitrary vector $\mathbf{x}$ according to the formula

$$I_{\boldsymbol{\beta}_\theta} = \left(\mathbf{x}\hat{\boldsymbol{\beta}}_{\theta-b}, \mathbf{x}\hat{\boldsymbol{\beta}}_{\theta+b}\right), \tag{4}$$

where $b = z_\gamma\sqrt{\frac{\mathbf{x}\mathbf{Q}^{-1}\mathbf{x}\cdot\theta(1-\theta)}{n}}$, $\mathbf{Q} = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i'\mathbf{x}_i$, $\gamma \in (0,1)$ – confidence probability (probability that the confidence interval will cover the true value), $z_\gamma = \Phi^{-1}(\gamma)$ – quantile of the standard normal distribution for probability γ, $\Phi^{-1}(\cdot)$ – function inverse to the standard normal distribution function.

Thus, to build a confidence interval for quantile regression estimates, we need to additionally estimate the quantile regression for probability levels $\theta \pm b$.

Getting data. The internal computer network of the enterprise was taken as the subject of the study. Using the SNMP protocol, data was collected that characterizes the operation of the network. Data on the functioning of the telecommunications network were recorded with an interval of 30 seconds. The following characteristics were determined for the analysis:

- number of lost packets;
- CPU frequency;
- processor temperature;
- bus voltage.

Data processing. All data were analyzed for missing observations and errors in values. For visual analysis, for each of the parameters, time series graphs were built (Fig. 1 and Fig. 2.).

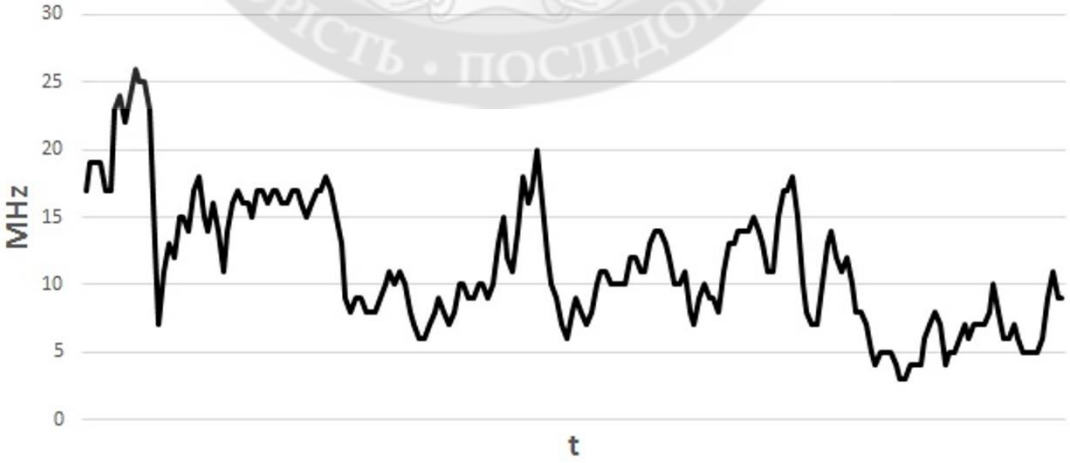

Figure - 2. Graph of processor load over time

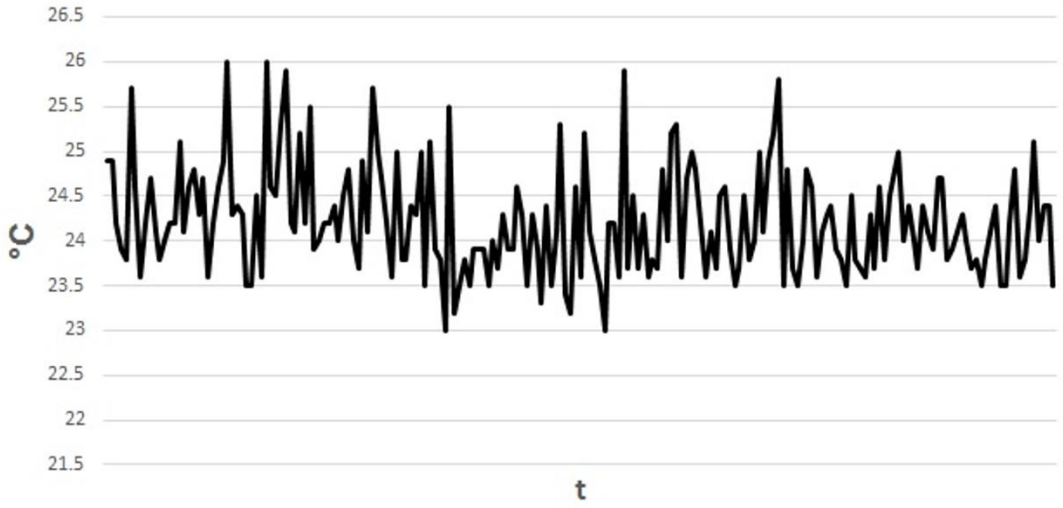

Figure - 3. Graph of CPU temperature over time

Conclusion

1. An analysis of the characteristics of the operation of digital radio electronic devices has shown that there is a pattern in the development of the values of these indicators over time. Moreover, there is no significant correlation between the characteristics, which means the independent nature of their behavior. Thus, the task of independent forecasting of the values of characteristics for the future makes it possible to prevent situations of overload of certain nodes of the telecommunication network. which was done in this work.

2. The possibility of using the quantile regression method to predict the values of parameters characterizing the operation of digital radio electronic devices has been proved.

3. The features of statistical monitoring of telecommunication networks are determined, namely: non-stationarity, periodicity (uneven loading of channels), non-linear influence of network operation characteristics on its efficiency.

4. Since the constructed quantile regression model could not fully explain the development of parameter values over time, the authors draw conclusions about the nonlinear dependencies of these indicators and set themselves the task of further research.

REFERENCES:

1. Khlaponin Y.I. Modern problems of creating complex real-time information and control systems / Khlaponin Yu.I., Nedaybida Yu.P., Kotova Yu.V. // Information protection. – 2012. – No. 4 (57). – P. 50-55.

2. Enyukov I.S. Statistical analysis and monitoring of scientific and educational Internet networks / I.S. Enyukov, I.V. Retinskaya; under. ed. A. N. Tikhonova. – M.: Finance and Statistics, 2004. – 320 p.

3. Khlaponin Yu.I. Modern problems of creating complex real-time information and control systems / Khlaponin Yu.I., Nedaybida Yu.P., Kotova Yu.V. // Information protection. – 2012. – No. 4 (57). – P. 50-55.

4. Ayvazyan S.A. Applied statistics: classification and dimensionality reduction / Ayvazyan S.A., Bukhstaber V.M., Enyukov I.S., Meshalkin L.D. Applied statistics: classification and dimensionality reduction. – M.: Finance and Statistics, 1989.

5. Andrews D.F. A Robust Method for Multiple Linear Regression // Technometrics. 1974. Vol. 16. P. 523–531.

6. Hogg R.V. Adaptive Robust Procedures: A Partial Review and Some Suggestions for Future Applications and Theory // Journal of the American Statistical Association. 1972. Vol. 43. P. 1041–1067.

7. Khlaponin Y.I. Establishment of neural measures in a statistical system of analysis and monitoring of telecommunication measures / Khlaponin Y.I., Zhirov G.B., Nikitchin O.M. - K.: Technological audit and production reserves, 2016. - P.35.

8. Koenker R., Bassett G., Jr. Regression Quantiles // Econometrica. 1978. Vol. 46. No.1. P.33–50.

9. Koenker R., Hallock K.F. Quantile Regression // Journal of Economic Perspectives. 2001. Vol. 15. No.4. P.143–156.

10. Huber P.J., Robust Statistics. 1981. New York: John Wiley and sons.
11. Karst O.J. Linear Curve Fitting Using Least Deviations // Journal of the American Statistical Association. 1958. Vol. 53. No. 281. P. 118–132.
12. Wagner H.M. Linear Programming Techniques Regression Analysis // Journal of the American Statistical Association. 1959. Vol. 54. No. 285. P.206–212.
13. Introduction to the theory of order statistics / Ed. A.E. Sarkhan and B.G. Greenberg. Per. from English Ed. AND I. Boyarsky. M.: Statistics, 1970.
14. Kenneth Q. Z., Stephen L. P. Direct Use of Regression Quantiles to Construct Confidence Sets in Linear Models // The Annals of Statistics. 1996. Vol. 24. No. 1. P. 287–306.
15. Serhey Lienkov, Genadiy Zhyrov, Ihor Pampukha, Ivan Chetverikov . Block Encryption Algorithm for Digital Information Using Open Keys for Selfgeneration of Closed Random Private Keys // 2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT), 18-20 Dec. 2019 Kyiv, Ukraine. Electronic ISBN: 978-1-7281-6144-0, Print on Demand(PoD) ISBN: 978-1-7281-6145-7, INSPEC Accession Number: 19452664 DOI: 10.1109/ATIT49449.2019.9030509.

д.т.н. проф. Хлапонін Ю.І. к.т.н. доц. Касім Н.Х., Тарасюк Д.М.

ПРОГНОЗУВАННЯ СТАНУ ТЕЛЕКОМУНІКАЦІЙНИХ МЕРЕЖ МЕТОДАМИ КВАНТИЛЬНОЇ ТА ЛОГІСТИЧНОЇ РЕГРЕСІЇ

У сучасному світі повсюдне поширення інформаційних технологій переплелись телекомунікаційні системи з усіма аспектами людського життя. Важко осягнути світ, де ви відключені від «Всесвітньої мережі» або не можете миттєво обмінюватися даними через заплутану мережу сучасних мобільних пристроїв. Життєвість підтримки зв'язку онлайн неможливо переоцінити, а забезпечення безперебійної роботи телекомунікаційних систем є першорядним. У цьому документі розглядається головне завдання прогнозування та керування продуктивністю цих мереж, використовуючи методи квантильної та логічної регресії. Наше дослідження використовує реальні дані з операцій телекомунікаційної мережі, щоб побудувати прогностичну модель, здатну заздалегідь передбачити стан мережі. Ця можливість прогнозування служить основою для інтелектуальних систем прийняття рішень, полегшуючи керування мережею в реальному часі. Впроваджуючи методи машинного навчання, зокрема квантильну регресію, ми досягаємо глибокого розуміння того, як різні фактори впливають на продуктивність мережі. Наше дослідження не обмежується лише прогнозуванням; він поширюється на сфери інтелектуальних систем прийняття рішень, де розуміння, отримане за допомогою регресійного аналізу, відіграє ключову роль. Ці інтелектуальні системи обладнані для прийняття керованих даними рішень щодо розподілу мережевих ресурсів, графіків технічного обслуговування та превентивного вирішення проблем. По суті, вони діють як зберігачі стабільності мережі, гарантуючи, що телекомунікаційні системи залишаються надійними та чуйними на вимоги сучасного суспільства, що постійно змінюються. Ця стаття проливає світло на незамінну роль регресійних методів у проактивному управлінні станом телекомунікаційних мереж. Використовуючи потужність машинного навчання та аналіз даних, ми прокладаємо шлях до майбутнього, де збої в роботі мережі зведені до мінімуму, а безперебійне з'єднання, на яке ми звикли покладатися, постійно буде присутнім у нашому житті.

Ключові слова: інформаційні технології, телекомунікаційна система, інтелектуальна система прийняття рішень, машинне навчання, квантильна регресія.