UDC 004.9:004.89
DOI: 10.15587/2706-5448.2022.255861
Article type «Reports on Research Projects»

Andrii Papa,
Yevgen Shemet,
Andrii Yarovyi,
Lyubov Vahovska

DEVELOPMENT OF INFORMATION TECHNOLOGY FOR ANALYZING THE CUSTOMER CHURN OF A TELECOMMUNICATION COMPANY

The object of research is the process of analyzing the customer churn of telecommunications companies based on machine learning methods. The existing problem is that, until recently, the process of customer churn was compensated by attracting new customers, but in the modern world, growth rates are constantly accelerating, the market is filled with a large number of competitors, which leads to a constant increase in customer requirements for products and services. In this regard, the process of attracting new customers becomes more costly and time-consuming, which in turn enhances the importance of maintaining an existing customer base.

The paper considers problematic aspects related to improving the accuracy of predicting the outflow of a company's customers through the use of machine learning methods. The conducted studies are based on the application of an approach implemented by combining the methods of decision trees and nearest neighbors. A positive result cannot be achieved by ignoring the existing resource constraints and requirements, which must be determined separately for each research case.

The relevance of the problem of analysis the outflow of customers for companies with many users is considered. A model for predicting the outflow of customers based on a combination of decision tree and nearest neighbor methods, which is used in the basis of the bagging method, is proposed. One of the features of this approach is the use of a test sample of normalized data. Accordingly, systems can use pre-known information, learn, acquire new knowledge, predict time series, perform classification, and in addition, they are quite obvious to the user. The prospect of choosing these methods is explained by the fact that they were used earlier in data analysis systems and provided sufficiently high-quality results.

The expediency and prospects of applying the proposed approach in the problem of analysis the outflow of customers of telecommunications companies are shown, as well as the design features of information technology and the results of software implementation.

Keywords: *machine learning, decision tree, nearest neighbour method, bagging, data analysis.*

Received date: 03.03.2022
Accepted date: 19.04.2022
Published date: 30.04.2022

© The Author(s) 2022
This is an open access article under the Creative Commons CC BY license

How to cite
Papa, A., Shemet, Y., Yarovyi, A., Vahovska, L. (2022). Development of information technology for analyzing the customer churn of a telecommunication company. Technology Audit and Production Reserves, 2 (2 (64)), 11–15. doi: http://doi.org/10.15587/2706-5448.2022.255861

1. Introduction

Currently, most companies that collect a large amount of data suitable for analysis use artificial intelligence methods, in particular machine learning and data mining [1, 2]. One of the popular examples of using machine learning in real life is the task of predicting customer churn. Telecommunications companies, banks, insurance companies and others are engaged in forecasting and managing customer churn. In a highly competitive environment, predicting customer churn in order to retain them is becoming one of the most important areas in modern business. As a rule, existing developments are based on the personal data of the client, as well as data on its activity in the company: the services and products that it uses, the history of transactional activity, the history of requests, information about purchases [3]. The data obtained are large arrays with structured and unstructured information, in which neural networks, data mining and machine learning methods are widely used to analyze and identify hidden patterns [4]. Despite the serious interest of analysts and scientists in the problem of preserving service consumers, based on the analysis of professional sources, one can state a certain limitation in the description of specific models, algorithms and software solutions. On the one hand, this is due to the significant specifics of the applied problem and the diversity of practical aspects of solving a specific problem in full for a specific service in a specific region. In addition, most economic problems have to be solved under conditions of initial information uncertainty [1, 4]. Thus, in order to solve the problem of analysis customer churn to retain existing users, in this study it is advisable to analyze machine

ISSN 2664-9969

learning methods [5], with the possibility of using them using the bagging method.

The object of research is the process of analyzing the outflow of customers of telecommunications companies based on machine learning methods.

The aim of research is to design an information technology containing a web server and an android application that will analyze and issue information in the form of a forecast of a possible outflow of customers of a given company based on input data.

2. Research methodology

For any business that provides goods or services, the customer base is important. New customers enter it, who actively use the services for some time and stop using the services after a while. This entire span is defined as the «customer life cycle» – a description of the stages that a customer goes through when it recognizes a product, makes a purchase decision, pays, uses and becomes a loyal consumer, and, ultimately, stops using products for certain reasons. Accordingly, the concept of «outflow» defines the final stage of the customer's life cycle, and for business, this means that the customer has ceased to make a profit and, in general, any benefit [3].

The outflow of customers is expressed in a reduction in the customer base and a decrease in revenue indicators [2]. Previously, the problem of customer churn was compensated by attracting new customers, but in the modern world, growth rates are constantly accelerating, the market is filled with a large number of competitors, which leads to a constant increase in customer requirements for products and services. In this regard, the process of attracting new customers becomes more costly and time-consuming, which in turn enhances the importance of maintaining an existing customer base.

The concept of customer churn is not a rigid concept and does not define the stage of termination. There are three main approaches to determining customer churn:

1. Churn is the refusal of customers to buy products and services of the company, the termination of service contracts by customers.

2. Churn is the termination of the use of the company's products or any of the services by the client.

3. Churn is a situation where a customer first actively uses a company's product or service and then reduces usage to a minimum.

Attracting new customers to combat churn is a more complex and resource-intensive procedure that requires a significant amount of funds for advertising, social networks and other customer acquisition channels than increasing the loyalty of existing customers. Also, the efficiency of operational work with existing customers is higher than with new ones, for several reasons [3]:

1. Existing customers are more loyal and willing to pay more if they are satisfied with the level of service.

2. Existing customers will purchase a new product or service with a probability of 70–80 %, for new customers this figure does not exceed 20–30 %.

3. It is not necessary to spend marketing budgets on customer acquisition in order to profit from them.

The first stage of work with the outflow of customers is the early identification of a group of people who are inclined to stop using services. Knowing in advance the possibility of customer churn, strategic decisions can be applied.

The main goal of customer churn analysis is to create a list of contracts (customers) that are likely to be terminated in the near future. There are different approaches to customer churn analysis. Most of them are based on machine learning methods, which show rather high efficiency in modern conditions.

Various mathematical models are used to predict customer churn, including logistic regression, decision trees, nearest neighbors, SVM, random forest [3, 4]. Since companies have limited ability to communicate with customers, it is important to determine the posterior probability of classification. Knowing the list of likely «refuse persons», the company can build an optimal strategy for holding promotions.

In this case, the prospect of creating a software product for predicting customer churn is a fairly high priority at the present time. After all, almost every business owner, built on working with clients and providing certain services, wants to be insured or at least warned about a possible outflow of customers. One often encounters the problem of the lack of software tools for predicting customer churn. Therefore, it is expedient to develop information technology to analyze the outflow of customers of a telecommunications company.

3. Research results and discussion

The development of the proposed architecture is based on the following system requirements and features:

– formation of an outflow forecast with an accuracy of at least 85 %;

– formation of an outflow forecast by at least two methods;

– availability of an API for generating a forecast of customer churn;

– the forecast formation time is less than 3 s.

In this case, the problem of predicting customer churn should be considered together with the problem of classification. That is, based on the known characteristics of the user, it is necessary to predict the membership in the group of those users who will go or stay. The classification problem is a supervised learning problem, i. e. required data sets: training and test sets. Statistical researchers have long used a method called «bootstrap sampling», which can be loosely translated as «bootstrap sampling variation». One of the embodiments of this idea in machine learning is bootstrap aggregating, or bagging for short, that is, combining results under different loads [5]. The idea of bagging is that in the absence of a large training sample, many random samples can be created with an initial simple substitution selection. Although elements in the samples may overlap or overlap in practice, the results of combining from many samples are still more accurate than just one initial one. The method is so called because it combines the prediction results of different classifiers trained on random subsets. Bagging turns out to be useful only in the case of different instabilities of classifiers, when small changes in the original sample lead to existing classification changes [6].

Let's consider in more detail the essence of the bagging method itself in the context of the task. Let there be a training sample X. Using bootstrap, let's generate samples X_1, ..., X_M from it. Now, on each sample, let's learn teach our own classifier $a_i(x)$. The final classifier will average the responses of all $a_i(x)$ (in the case of classification, this corresponds to voting). The visualization of this circuit is shown in Fig. 1.

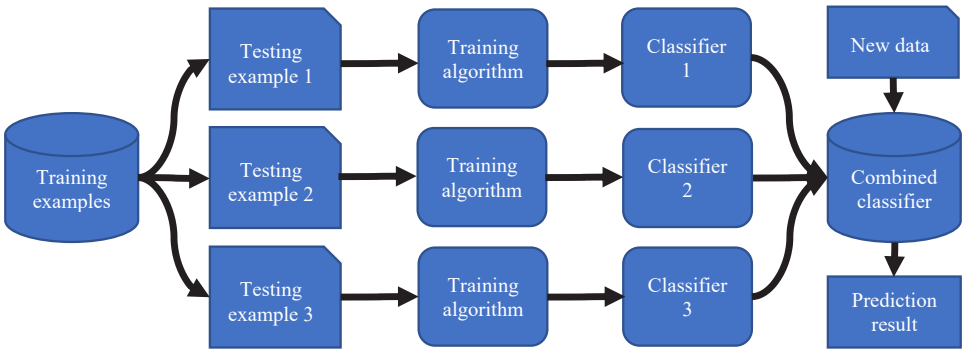

Fig. 1. Scheme of how the composition of methods works

Bagging allows reducing the variance when training the classifier, reducing the value showing the degree of difference in the error, provided that the model is trained using different data, or, in other words, is a warning link for retraining. The effectiveness of the method is achieved due to the fact that the basic algorithms that have been trained on different subsamples turn out to be quite different, and their errors are mutually compensated during voting, and also due to the fact that exception objects may not fall into some training subsamples.

From the available scientific studies, it can be concluded that the error of dispersion is much smaller with bagging, as indicated above. Bagging is effective on small samples, when the exclusion of even a small part of training objects leads to the construction of significantly different base classifiers. In the conditions of large samples, as a rule, subsamples of a significantly smaller length are generated [7].

When designing the client part of the information technology for analyzing customer churn, a block diagram was developed. First of all, it is possible to pay attention to the fact that the construction of information technology for analyzing customer churn is a complex of individual modules. A module in this context is a component of the whole system, defined by a certain property. That is why the structure of any information system can be represented by a set of subsystems.

The developed architecture of information technology for analysis the outflow of customers of a telecom company is shown in Fig. 2. Incoming user data is represented by records in the database. The forecasting module then uses this data to create forecasting models. The module includes generated models: a decision tree model and a kNN (nearest neighbors) model. This information development provides data in API format for generating customer churn forecasts. For this, a server module has been developed that interacts with the forecast generation module. The server module is also responsible for validating incoming values from information technology clients.

The information technology client, represented by the Android application, provides an interface for entering information about a new client, interacts with the server using HTTP requests.

To develop the client part of a software product that provides an analysis of the outflow of customers of a telecom company, it is necessary to determine the structural elements that implement the functions necessary to solve the tasks.

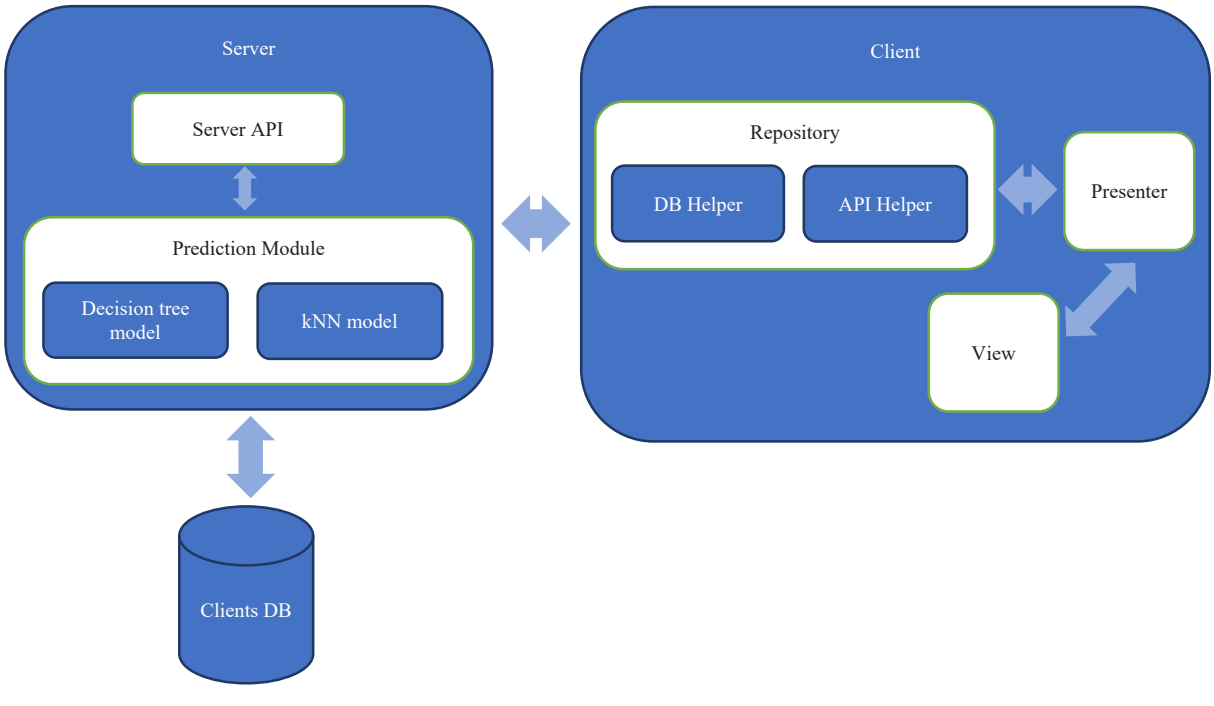

Fig. 2. Architecture of information technology for analysis the outflow of customers of a telecommunications company

ISSN 2664-9969

To ensure content protection, it is necessary to develop a login module that implements user authorization and registration. User data, their reports, files, rules used to analyze the outflow of customers must be stored in the database. Before event prediction and analysis, it is necessary to enter data into the system (in the form of a questionnaire), which will be processed in the data analysis module. The decision module makes recommendations for choosing one of the two events. The customization module involves making changes to the data and the system.

Fig. 3 shows an IDEF0 diagram of the first level of software decomposition that analyzes the outflow of customers of a telecom company. Input data are user data, client characteristics, environment. Purpose: This process should be modeled to demonstrate the operation of current (AS-IS) processes in the designed information technology in the form of a model representing hierarchically ordered and interconnected diagrams illustrating the connection of the system with external entities and the interconnection of the internal processes of the system. The user gains an understanding of the organization of the system as in as a whole, and its individual functional blocks. Viewpoint: When building the model, the system was considered from the point of view of its users. Definition: The model is created to illustrate the operation of the system at different levels of decomposition. Scope: General management of system data: acquisition, processing and storage.

Also, for a visual representation of the structure of information technology, a DFD model has been developed. Data flow diagram (Data Flow Diagram) design model, graphical representation of data flows in an information system. The data flow diagram can be used to visualize data processing processes (structural design) [8].

It is common for a developer to develop a context level data flow diagram first, which will show how the system interacts with external modules. This diagram is further refined by detailing the processes and data flows to show the extensive system under development. Data flow diagrams contain four types of graphic elements [8]:

1) processes – represent the transformation of data within the described system;

2) data storage (repositories);

3) entities external to the system;

4) data flows between elements of the three previous types.

Information sources (external entities) generate information flows (data flows) that transfer information to subsystems or processes. Those, in turn, transform information and generate new flows that transfer information to other processes or subsystems, data accumulators or external entities – consumers of information [9].

This methodology (the Gane/Sarson methodology) is based on the construction of an IS model designed or actually existing. According to the methodology, the system model is defined as a hierarchy of data flow diagrams (DFD) that describe the asynchronous process of information transformation from its input into the system to its issuance to the user. Diagrams of the upper levels of the hierarchy (context diagrams) define the main processes or IS subsystems with external inputs and outputs. They are detailed by lower level diagrams [10]. The developed DFD model is shown in Fig. 4.

The obtained results of the analysis of machine learning methods for predicting the loss of customers show the feasibility and prospects of using the chosen ones. Therefore, an information technology for analyzing customer churn is proposed, which is distinguished by a combination of decision tree and nearest neighbor (kNN) methods, which is used in the basis of the bagging method for analyzing customer churn, which improved the accuracy of customer churn forecasting.

In the course of the study, an information technology for analyzing customer churn was developed and programmatically implemented, as well as testing it and analyzing the results obtained. Testing showed full compliance of information technology with the analysis of customer churn to the tasks set, namely, the analysis process was completed in less than 2 seconds, the accuracy of forming a customer churn forecast was more than 85 %. However, there are limitations that must be adhered to in order to obtain the current result. Such restrictions are: the volume of the training sample is not less than 1000 records, the need for a preliminary training procedure, which can take from 2 hours or more. Also, certain technical limitations should be added: a stable Internet connection, a relatively limited number of requests to the server, a Unix-like operating system, a sufficient amount of system RAM (at least 8 GB). When comparing the developed product with analogues, the best results were obtained by using a combination of machine learning methods and their combination at the heart of the bagging method.

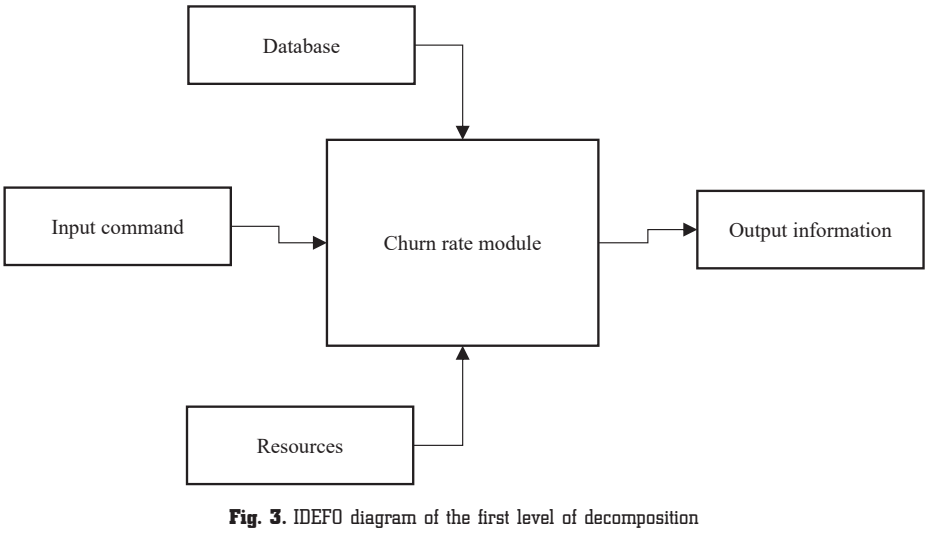

Fig. 3. IDEF0 diagram of the first level of decomposition

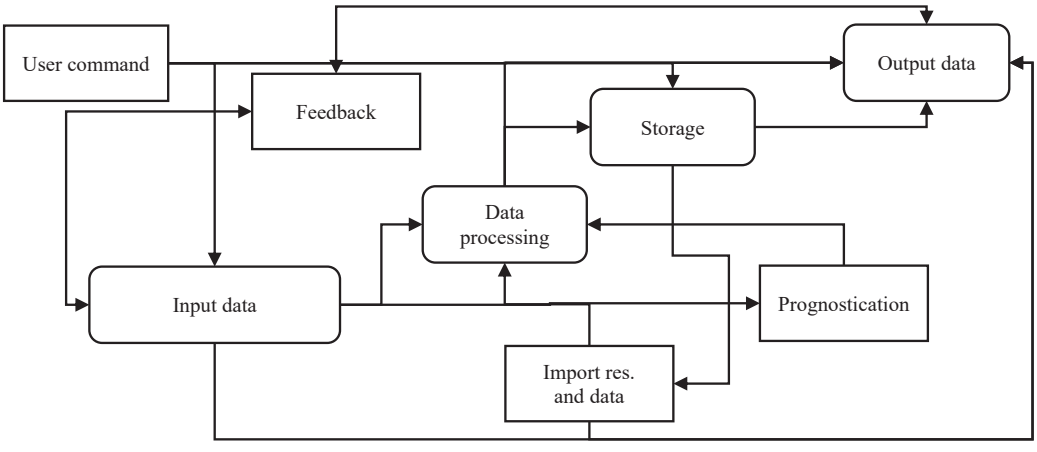

Fig. 4. DFD diagram of information technology for analyzing the outflow of customers of a telecom company

In the future, it is planned to explore other combinations of machine learning methods, as well as artificial intelligence methods, to implement improved analysis and improve the accuracy of customer churn forecasts. Let's also plan to improve the machine learning procedure in order to increase performance.

4. Conclusions

The analysis of machine learning methods for predicting customer churn has been carried out, the results of which confirmed the feasibility and prospects of using combinations of several methods to solve the problem of predicting customer churn. A telecom company's client churn forecasting model has been developed, which differs from the well-known combination of decision trees and nearest neighbors (kNN) methods used in the basis of the bagging method, which improved the accuracy of forecasting the churn of a particular client. The client and server parts of the information technology for analyzing the outflow of customers of a telecom company have been designed and implemented in software, and its testing has been successfully carried out.

References

1. Papa, A. A., Yarovyi, A. A., Prozor, O. P. (2019). *Information technology analysis of the outflow of customers by telecom company*. Available at: https://conferences.vntu.edu.ua/index.php/all-fitki/all-fitki-2019/paper/view/7324
2. Papa, A., Shemet, Y., Yarovyi, A. (2021). Analysis of fuzzy logic methods for forecasting customer churn. *Technology Audit and Production Reserves, 1 (2 (57))*, 12–14. doi: http://doi.org/10.15587/2706-5448.2021.225285
3. Huang, B., Kechadi, M. T., Buckley, B. (2012). Customer churn prediction in telecommunications. *Expert Systems with Applications, 39 (1)*, 1414–1425. doi: http://doi.org/10.1016/j.eswa.2011.08.024
4. Tsai, C.-F., Lu, Y.-H. (2009). Customer churn prediction by hybrid neural networks. *Expert Systems with Applications, 36 (10)*, 12547–12553. doi: http://doi.org/10.1016/j.eswa.2009.05.032
5. Mahesh, B. (2020). Machine learning algorithms – a review. *International Journal of Science and Research, 9*, 381–386.
6. Bühlmann, P., Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical science, 22 (4)*, 477–505. doi: http://doi.org/10.1214/07-sts242
7. Liang, G., Zhang, C. (2010). Empirical study of bagging predictors on medical data. *Conferences in Research and Practice in Information Technology Series*. Ballarat.
8. Ibrahim, R., Yen, S. Y., Pahat, B. (2011). *A Formal Model for Data Flow Diagram Rules 1*. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.206.5214
9. Aleryani, A. Y. (2016). Comparative study between data flow diagram and use case diagram. *International Journal of Scientific and Research Publications, 6 (3)*, 124–126.
10. Kovalenko, O. S., Dobrovska, L. M. (2020). *Proektuvannia informatsiinykh system: Zahalni pytannia teorii proektuvannia IS*. Kyiv, 192.

✉***Andrii Papa,*** *Postgraduate Student, Department of Computer Science, Vinnytsia National Technical University, Vinnytsia, Ukraine, e-mail: papa.andriy@gmail.com, ORCID: https://orcid.org/0000-0002-7753-8576*

Yevgen Shemet, *Postgraduate Student, Department of Computer Science, Vinnytsia National Technical University, Vinnytsia, Ukraine, ORCID: https://orcid.org/0000-0001-5067-1900*

Andrii Yarovyi, *Doctor of Technical Sciences, Professor, Head of Department of Computer Science, Vinnytsia National Technical University, Vinnytsia, Ukraine, ORCID: https://orcid.org/0000-0002-6668-2425*

Lyubov Vahovska, *Assistant, Department of Computer Science, Vinnytsia National Technical University, Vinnytsia, Ukraine, ORCID: https://orcid.org/0000-0002-4865-6514*

✉*Corresponding author*