Evaluation of the quality of functioning of a telecommunications company based on Data Mining technology

Globa Larysa, Novogrudska Rina, Moroz Anastasia
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
Kyiv, Ukraine
lgloba@its.kpi.ua, rinan@ukr.net

***Abstract*—The paper presents an approach to assessing the quality of a telecommunications company, which can be used to predict the churn of subscribers from a telecom operator. The originality of the approach lies in the use of such mathematical methods that make it possible to determine the main set of parameters due to which specific subscribers are inclined to change their mobile operator. The approach is based on data mining, namely the method of cluster analysis, which most fully allows assessing the quality of telecommunications companies functioning. Mathematical modeling of predicting customer churn using methods of association rules, decision trees and bagging has been carried out. The parameters characterizing the interaction of a mobile operator with end users are described. The parameters that have the greatest influence on the client's decision to refuse the services of a mobile operator have been determined.**

***Keywords*—Telecom operator, Quality assessment, Customer churn, Data Mining, Modeling.**

I. Introduction

The billing model, in which the collection and processing of data was carried out by one specialized automated billing system that perfors only finensial calculations, does not satisfy the needs of the telecom operator. It is being replaced by models where a separate system is involved in collecting, processing and preparing data. The main reasons for the development of such a system are:

- The market demands from the operator a variety of offered tariff plans and service packages, quick commissioning of services, and therefore carrying out various analyzes of the accumulated data and according to various criteria.
- Telecom operators are forced to provide services on dissimilar equipment, which significantly complicates the collection of statistics through a variety of accounting data formats.
- Large telecom operators are expanding geographically to include regional companies. This leads to a sharp increase in heterogeneous equipment and the volume of accounting data.

At the same time, there is a need for centralized accounting of the provided communication services in the conditions of operation of several different billing systems in the regions. In the course of their development, operators developed their own data preprocessing systems. The complexity of the operation of such systems lies in the need for their constant revision in connection with the emergence of new services, new equipment, and changes in the data formats coming from the equipment. In this case, the operator is forced to use the resources of a team of highly qualified programmers to constantly improve the system. A standard situation is when the data preprocessing is carried out by various systems at a telecom provider company.

With the modern development of telecommunications, systems that calculate the consumed services by users provided on the basis of one or another telecommunications equipment have received the same development. Such systems are usually called billing systems [1]. Their main purpose is to display the number of services consumed by the subscriber and write off funds in accordance with the cost per service unit. Based on this, it is possible to give a definition: billing system (BS) is a software complex that records the volume of services consumed by subscribers, calculates and debits funds, according to tariffs. In terms of relationships with content providers, the billing system must provide the following capabilities: firstly, it must guarantee the transmission of information to partners about the use of their services by specific subscribers, and secondly, bill the end consumer for using the service (or the content provider independently issues an invoice).

In the context of the rapid development of the communication network, replacement and addition of switching equipment, the provision of new types of services, it becomes expedient to switch to an industrial solution of data preprocessing problems. Products of this class have long been successfully used by large operators in countries with developed communication infrastructure. They are called Mediation systems. Mediation systems represent the level between the network infrastructure and OSS / BSS (Operation Support System / Business Support System) systems. The main purpose of Mediation systems is to transform data received from network

elements into information that can be interpreted by the billing system and other business systems of the operator. Also, such systems are called previous billing systems.

In the context of sustainable development of telecommunications, as well as the presence of a sufficient number of provider companies in the telecom services market, important tasks are to assess the quality of telecom operators' functioning and develop a strategy to prevent subscriber churn. However, for this it is necessary to increase the level of intellectualization of OSS / BSS systems and to include in their structure tools for processing and analyzing large volumes of accumulated data.

II. State of art and background

Currently, the usual methods of increasing the loyalty of old customers and attracting new customers (mass advertising, traditional marketing, low prices) do not produce the desired positive result. That is why concepts that allow personalized sales of goods and services are becoming a priority all over the world.

An example is Customer Relationship Management (CRM) – customer relationship management [2]. CRM is not a technology or software product. This is a business strategy based on a customer-oriented [3] approach, that is, we can say that the main task of CRM is to increase and improve the efficiency of business processes, directly aimed at attracting new and retaining existing clients [4]. CRM systems are most effective in those areas of business where it is possible to accumulate a large amount of useful information about each client. Here, the CRM strategy is used primarily to combat customer churn. In telecommunications, this term refers to the process of enticing customers from one telecom operator to another, or simply the outflow of customers from the operator [5]. The annual rate of customer churn reaches 25 – 30%. Operators who have this indicator is maximum, will not be able to get a return on investment in new subscribers, since it takes about three years to get back the funds spent on replacing each lost client with a new one, that is, it takes about three years to acquire clients [6, 7].

Currently, OSS / BSS have gained active development to support the operational and business activities of telecom operators, with the help of full or partial automation of these activities. Operational activities include processes that interact mainly with network equipment and the network, such as: accounting and planning of network resources, management and provision of services, management of quality characteristics of services. Business activities include customer-centric processes such as processing and invoicing, collecting payments, proposing new products, and much more [8].

Thus, both OSS / BSS systems and CRM systems, although they allow to some extent process big data of telecom operators, it is necessary to integrate them and supplement them with methods of deeper analysis and structuring of such data to solve the problem of assessing the effectiveness of the telecom operator's functioning, and also predicting the churn of customers from the operator.

A. Data Mining in telecommunications

Today, the Data Mining solution [4] is widely used by telecommunications operators, network operators and Internet providers in order to improve the processing of unstructured large amounts of data for making better decisions in their business. Analytics enables telecommunications providers to significantly improve the cost-effectiveness of their service delivery.

Data Mining models are positioned on a dataset with known results and are used to predict results from other datasets. These are classification models (they describe the rules by which the description of an object can be attributed to one of the classes) and sequence models (they describe functions by which it is possible to predict the change in continuous numerical parameters).

In telecommunications, it is very important to identify categories of customers with similar stereotypes for using services, as a result, a pricing policy is being developed; analysis of failures; prediction of peak loads. Telecom systems generate an extremely large amount of data that needs to be processed and analyzed. Thus, data mining is becoming a very important and widely implemented part of business processes in the systems of a telecom operator [9].

Data Mining methods make it possible to effectively solve the problems of structural engineering design of innovative technical systems in telecommunications [2]. These methods have much in common with methods for solving problems of classification, diagnostics and pattern recognition. Nevertheless, one of their main distinguishing features is the function of interpreting the patterns that form the basic rules for including objects in equivalence classes. Therefore, such methods are becoming more common today. Developing intelligent data processing for telecommunications companies is essential to:

- Reducing the computational complexity of big data processing methods for the operator to provide services to a subscriber with a given quality of service;
- Predicting the risks that may arise from the operation of the telecommunications system;
- To be able to identify faults in the system and find out the cause of their occurrence.

Some of the existing algorithms can be adapted to compute large distributed information arrays. At the same time, serious difficulties can arise in the visual presentation of the results – due to the huge amount of information entering the input, the number of different reports at the output increases dramatically. For their convenience, new mathematical methods are needed, which

are fundamentally different from the report generators used for traditional storage. In this regard, it is relevant to use just such mathematical methods that make it possible to move from obtaining data to obtaining information, and from it to obtaining knowledge about the process and patterns of providing services to subscribers and, thus, increasing the efficiency of managing this process.

III. The method of clustering big data of a telecom operator

In data mining tasks, using cluster analysis, a complex construction of data for classification is created, patterns are identified, hypotheses are formed and tested, etc. In addition, cluster analysis is often used to identify data that "stands out" from others, since such data correspond to points located at a distance from any cluster. Cluster analysis is used to compress and summarize data. A method for clustering big data of a telecom operator is proposed, the use of which allows assessing the quality of functioning of telecommunication companies [10]. The method uses decision trees, association rules, and bagging to predict customer churn from a carrier. The method proposed in the study allows you to go through all the stages of preparation, processing and analysis of big data of a telecom operator. The method includes next stages:

- Subject domain analysis.
- Problem statement.
- Data preparation.
- Model development.
- Model checking.
- Model selection.
- Model preparation.
- Model updating.

The following systems are responsible for the first three stages: analysis of the subject area, statement of the problem, preparation of data: billing, OSS / BSS. The following steps are performed by the CRM system.

All of the above stages together give a high-quality result in solving the problem of customer churn in the telecommunications sector. This question is very important for doing business and determining the new course of the company.

The first thing to do is to analyze the telecommunications structure. Data that will be processed in the future are collected from network equipment. Further, in order to solve the problem of determining the cause of customer churn, it is necessary to identify which factors influence this, which data components are of the greatest importance.

After collecting the data, they need to be structured. To do this, the "raw" data must go through the following steps: verification, correction, consolidation, filtration, separation, routing, representation, distribution.

The next stage is modeling.

IV. Modeling and predicting the churn of subscribers of a telecom operator

The input data for the modeling were real customer data obtained from one of the leading telecom operators in Ukraine (Fig. 1). We will call such data "raw". There are spaces in the data table, which is invalid for analysis. Gaps can introduce additional prediction errors.

ABON_CODE	STATUS	COUNT_DAYS_OVER_1MB	COUNT_DAYS_OVER_5MB	DUAL_SIM_PROBABILITY	SIM_PRIORITY	OBLAST	CITY
1	0			High	SECOND	Donets'ka	Rodynske
2	0			Very high	SECOND	Kharkivs'ka	Izyum
3	0			Very high	SECOND	L'vivs'ka	Novoiavorivs'k
4	0			High	SECOND	Kharkivs'ka	Kharkiv
5	0	10	9	Very high	SECOND	Kharkivs'ka	Bezlesdubrivs
6	0	23	19	Very high	FIRST	Donets'ka	Mykhylka
7	0	27	23	Low	FIRST	Kyivs'ka	Kyiv
8	0			Very high	SECOND	Kharkivs'ka	Kharkiv
9	0			Very high	FIRST	Kharkivs'ka	Dar Kolodiaz'
10	0			High	SECOND	Kharkivs'ka	Zavhorodnie
11	0			Very high	SECOND	Donets'ka	Artemivs'k
12	0			Very high	SECOND	L'vivs'ka	Kolodostil
13	0			UNDEF	UNDEF	L'vivs'ka	Khodoriv
14	0			Low	FIRST	Odes'ka	Shyrioka Balka
15	0			UNDEF	UNDEF	Kyivs'ka	Kyiv
16	0			Very low	FIRST	L'vivs'ka	Boryslav
17	0			Low	FIRST	Ternopil's'ka	Ternopil'
18	0			High	SECOND	Donets'ka	Mykitivka
19	0			High	SECOND	Kyivs'ka	Kyiv
20	0			High	FIRST	Kharkivs'ka	Artil'ne
21	0	6	4	Very high	SECOND	Ternopil's'ka	Romanove Sele
22	0	2		Very high	SECOND	Zhytomyrs'ka	Dmytychve
23	0	4	3	Very low	FIRST	Zakarpats'ka	Chop
24	0			Very high	SECOND	L'vivs'ka	Hirne
25	0			Very high	SECOND	Rivnens'ka	Sarny
26	0	8	7	Low	FIRST	Rivnens'ka	Zdolbuniv
27	0			Very low	FIRST	L'vivs'ka	L'viv
28	0			Very high	SECOND	Mykolaivs'ka	Mykolaiv

Figure 1. Initial modeling data.

For the correct execution of the process of analyzing the dataset, you need to fill in the gaps that are present in the input table. To do this, use the search function for gaps and fill them with zeros (Fig. 2). Thus, when simulating, we get a high probability of truthful results. As a result, we will receive "preliminary prepared data".

	STATUS	COUNT_DAYS_OVER_1MB	...	INET_SLOPE	REFILL_SLOPE
0	0	0.0	...	0.000000	0.000000e+00
1	0	0.0	...	0.022223	-4.350000e-36
2	0	0.0	...	0.046156	3.710257e-02
3	0	0.0	...	-0.026511	3.838383e-02

Figure 2. Pre-prepared data for modeling.

When modeling customer behavior in order to predict subscriber churn the random forest method was used, with the help of which the modeling is carried out. The random forest is one example of ensemble classifiers. The modeling was performed in the Python programming language. In the course of the study, a sample of subscribers inclined to outflow from the given probabilities was determined (Table 1)

Table I
SUBSCRIBERS WHO ARE PRONE TO CHURN

$N^{\underline{o}}$	prob_true
1977	0.9333
2696	0.9167
2708	0.9233
2924	0.9041

The work uses Data Mining methods: associative rules, decision trees and bagging in order to increase the efficiency of prediction by increasing the accuracy of the probabilities of subscriber churn.

It was found that the bagging method allows you to get more accurate results, by 7% compared to the method

of association rules and up to 1% better than the value of the metric of the decision tree method, the accuracy of the results obtained is 78.84%.

Also, in the process of modeling, the problem of identifying the factors that most affect the decision of the subscriber to switch to the services of another telecommunications company was solved. The modeling revealed the following parameters that most affect the churn of customers, they are presented in table. 2. These parameters are:

- duration of using the services of the telecommunications company;
- the number of days during which the subscriber does not use the services;
- average value of all days during which the subscriber does not use services;
- average value of all active days outgoing calls to numbers of other mobile operators.

Table II
IMPACT OF PARAMETERS ON THE CUSTOMER CHURN PROCESS

N^o	importance	labels
3 2	0.075648	Duration of using the services of the telecommunications company
2 3	0.068619	The number of days during which the subscriber does not use the services
2 1	0.060132	Average value of all days during which the subscriber does not use services
2 2	0.055328	Average of all active days
1 9	0.045268	Outgoing calls to numbers of other mobile operators

The simulation revealed the subscribers who are most susceptible to the transition to the services of another telecommunications company. Table 3. shows a sample of subscribers prone to churn from a mobile operator. The churn rate percentage is shown next to each subscriber. It can be concluded that, based on the consideration of these data, it is possible to influence individual users and provide each of them with those services in which the subscriber is most interested. This way you can prevent customer churn.

Table III
SUBSCRIBERS WHO ARE PRONE TO CHURN

N^o	prob_true
13788	0.99256
13859	0.99121
17185	0.98286
11595	0.95789
10842	0.94865

CONCLUSION

A telecom operator collects large amounts of data in the course of its work; for full processing, this data must be structured. Large amounts of data are processed using systems: OSS / BSS, billing, CRM. Correct storage, structuring and analysis of this data in aggregate with high accuracy allows predicting the behavior of the subscriber and his use of the operator's services, as well as the possibility of the subscriber's transition to another operator.

Mathematical modeling for predicting customer churn was performed using the methods of association rules, decision trees and bagging. Such methods allowes to get more accurate results, by 7% compared to the method of association rules and up to 1% better than the value of the metric of the decision tree method, the accuracy of the results obtained is 78.84%. Also, as a result of the study, patterns and factors were identified that most affect the subscriber's decision to refuse the services of a telecom operator.

The use of data obtained in the course of research is especially effective at the stage of concluding an agreement with a client, which allows you to build relationships with a client in the most beneficial way for the company.

REFERENCES

[1] Nechiporuk, DV, Features of Data Mining Technology, Don State Technical University, Rostov-on-Don, Russian Federation.

[2] Skorbota, V. S., "How Ukrainians choose a mobile operator," (2017) [Online]. https://biz.nv.ua/experts/skorbota/kak-ukraintsy-vybirajut-mobilnogo-operatora-1930612.html.

[3] Dyakonov PV, Handbook on the use of the PC MATLAB system, Moscow: "Fiz-Matlit", (1993).

[4] Palmov, SV Comparison of the capabilities of various methods of Data Mining technology in the analysis of personal traffic. // XII Russian scientific conference of professors and teaching staff, researchers and graduate students, Samara, PDATU, (2005), thesis, pp. 285–287.

[5] Sasaki, Y., The truth of the F-measure, 26th October, (2007)

[6] Scholl, F., Deep Learning in R, St. Petersburg: Peter Publishing House, (2018).

[7] Kaftannikov, IL, Parasich, AV, Features of the use of decision trees in problems South Ural State University, Chelyabinsk.

[8] ALEXANDER, Schill; LARYSA, Globa; TETIANA, Kot. Applying business process modeling method when telecommunication services development. In: 2011 21st International Crimean Conference "Microwave & Telecommunication Technology". IEEE, 2011. pp. 457–458.

[9] Globa, L.S., Novogrudska, R.L., Koval, A.V.: Ontology Model of Telecom Operator Big Data. Proceedings of IEEE International Black Sea Conference on Communication and Networking (BlackSeaCom), 1–5 (2018). doi:10.1109/BlackSeaCom.2018.8433710.

[10] Y Buhaienko, LS Globa, A Liashenko, M Grebinechenko. Analysis of clustering algorithms for use in the universal data processing system. In: Open Semantic Technologies for Intelligent Systems. Research Papers Collection, Issue 4, Belarusian State University of Informatics and Radioelectronics, Minsk, 2020, pp. 101–104

Оценка качества функционирования телекомуникационной компании на основе технологии Data Mining

Л. Глоба, Р. Новогрудская, А. Мороз

В статье представлен подход к оценке качества функционирования телекомуникационной компании, который может быть исспользован для предсказания оттока абонентов от оператора связи. Оригинальность подхода заключается в использовании таких математических методов, которые позволяют определить основной набор параметров, из-за которых конкретные абоненты склонны к смене мобильного оператора. Подход основан на data mining, а именно методе кластерного анализа, который в наиболее полной мере позволяет оценить качество функционирования компаний телекоммуникационной связи. Проведено математическое моделирование предсказания оттока клиентов с помощью методов ассоциативных правил, деревья решений и bagging. Описаны параметры, характеризующие взаимодействие оператора мобильной связи с конечными абонентами. Определены параметры, оказывающие наибольшее влияние на решение клиента об отказе от услуг мобильного оператора.

Received 14.05.2021