Approach to Prediction of Mobile Operators Subscribers Churn

Andrii Baria, Larysa Globa, Anastasiia Moroz
National Technical University of Ukraine ”Igor Sikorsky Kyiv Polytechnic Institute”
Kyiv, Ukraine
baria.andrii@gmail.com, lgloba@its.kpi.ua, a.moroz2708@gmail.com

***Abstract*—This article presents an approach to the description of machine learning methods for predicting the outflow of telecom operator subscribers. Describes the parameters characterizing the interaction of the mobile operator with end users. The parameters that have the greatest influence on the client's decision to refuse the services of a mobile operator have been determined. The originality of the approach lies in the use of such mathematical methods that allow you to determine the main set of parameters, due to which specific subscribers are prone to changing the mobile operator. The proposed approach allows you to organize a system in which it is possible to determine the main parameters characterizing the tendency of customers to outflow and acting on them using various methods to try to increase subscriber loyalty. A comparative analysis of the results obtained using the analyzed logistic regression methods, Bootstrap aggregating and random forest showed that the spread of prediction errors does not exceed 6%. However, the advantage of the random forest method is the ability to determine the set of parameters that make the greatest contribution to making decisions by the subscriber to change the mobile operator. Therefore, for analyzing customer loyalty, a random forest method can be recommend, which showed on the test sample an improvement in the accuracy of the predictions in the sample to 6-7%.**

***Keywords*—telecom operator, churn, machine learning, random forest, prediction mathematical methods.**

I. Introduction

Constantly developing mobile market creates great competition, where subscribers tend to constantly look for the most favorable conditions for the provision of services by mobile operators. In this regard, operators need to constantly offer the most relevant services for each subscriber in order to keep this subscriber in their communication network. This is because the cost of retaining the customer is significantly less than attracting new customers. [1] Telecom operators are constantly analyzing the parameters characterizing the use of services by the end subscriber to identify those factors that have the greatest influence on the decision of subscribers to abandon the use of services, as well as to identify those subscribers who in the future will be at risk. [2] The task is complicate by the presence of a huge amount of information collected, which cannot be process by the ”old” methods of information analysis.

To solve such problems can be used different methods of machine learning, which do not always give good results due to the complexity of the data. In this regard, the process of building models is reduce to choosing the most appropriate machine learning method for a particular case with a specific data set, which is a nontrivial task that requires a specialized approach.

To assess the effectiveness of the use of mathematical methods, an experiment was conduct using mathematical modeling on the data array of one of the major Ukrainian mobile operators in order to answer the question ”why do subscribers leave?”

The experiment is divide into 2 stages. The first stage is to determine the reasons why subscribers decide to stop using this or that service. The second stage of work consists in compiling a list of subscribers who are prone to care, indicating the percentage of probability and the parameters that have had the greatest influence on decision making.

The structure of the article: Part 2 contains an analysis of work on the outflow of subscribers from companies providing various services. Part 3 provides an overview and analysis of data provided by one of the major mobile operators. Part 4 describes the metrics for the proposed assessment method with reference to the parameters characterizing subscribers who have the greatest influence on the process of a classifier designing. Part 5 presents the results of the prediction of outflow of subscribers from the mobile operator. Part 6 includes conclusions and recommendations for further work.

II. STATE OF THE ART AND BACKGROUND

In [3] and [4], they talk about solving a similar problem for companies providing Internet services. For the analysis, such algorithms as logistic regression method, decision tree, and neural network are considered. The best results were obtain on a logistic regression model with a prediction accuracy of 89% and a sensitivity of 91%.

Alfa-Bank uses the Oracle Exadata storage and processing platform, the Oracle Big data Appliance and the Hadoop framework to analyze social networks and user behavior, assess creditworthiness, forecast customer outflow, personalize content and secondary sales [5], [6].

For the segmentation and management of customer churn, financial reporting, analysis of feedback on social networks and forums, VTB24 uses Teradata, SAS Visual Analytics and SAS Marketing Optimizer [7], [8].

All these methods make it possible to obtain a forecast of future trends on the existing data sets, but the definition of the set of parameters that made the greatest contribution to the current situation remains not fully resolved.

Research results show that in order to get an adequate solution to a specific problem, it is necessary to take into account the total metrics, the way of collecting data, their initial structure, ways of processing data and selecting the best parameters for models. All this requires an analysis of mathematical methods and models for solving such problems from the point of view of accuracy, completeness and correctness of the results obtained. Therefore, the initial stage is to correctly select the necessary methods and build the appropriate models of data structures that allow you to get the most adequate solution to a specific task on a specific data set. The analysis of the solution of the problem of predicting the outflow of subscribers from mobile operators allowed us to identify the most promising mathematical methods that demonstrate the best results in solving such problems: decision trees, random forest, k-means, naive bays classifier, bootstrap aggregating.

Analysis of the methods of machine learning is carried out subject to the adoption of a compactness hypothesis, which assumes that classes form compactly localized subsets in the space of objects. In general, such a hypothesis is fuzzy, since all classes of recognition a priori intersect in the space of signs. In this case, for the formalization of the concept of ”similarity”, the function of distance or the metric d (x, y) in the N-dimensional space of objects is introduced.

In practical tasks, for the purpose of obtaining the values of the function of correspondence, a vector of signs of a functional state, which consists of both continuous (quantitative) and discrete (categorical) attributes taking their values from a finite disordered set, are used. With the data within the nominal scale, in which categorical signs are measured, no arithmetic operations can be performed, since all types of numerical processing relate to the ordering of objects in each class. Bringing categorical primary attributes to quantitative secondary by simply numbering the values of the primary attributes rarely results in satisfactory results, since the algorithms will take into account ordering that has no meaning, so processing mixed-type data causes some difficulties and is undesirable.

The disadvantage of algorithms that use a remote metric is to ignore additional information that is described by statistics of qualitative characteristics. For example, for a categorical sign it is possible to calculate its frequency (number of observations) and fashion (the value that has the highest frequency).

Metric algorithms perform a local sampling approximation, in which calculations are delayed until a known input object becomes known. Metric algorithms refer to methods of lazy learning (lazy learning).

The fundamentals of eager learning methods that make global sampling approximation were embodied in the theory of multivariate statistical analysis and decision-making theory.

The essence of statistical teaching methods is to restore a separate function by minimizing the average risk of false decision making. Statistical methods allow constructing deciding rules in the cases of crossing the recognition classes, which takes place in practical problems of control and management of weakly formalized processes.

The main drawbacks of statistical methods that restrict their use in practice are the need for large volumes of statistics to approximate the probability density distribution function, to fulfill rigid conditions for ensuring statistical stability and homogeneity and high sensitivity to the representativeness of the training samples.

The main disadvantage of the SVM (support vector machine) method is the limitation of its use for the tasks of analyzing the outflow of subscribers due to the algorithm's model due to the ignoring of the a priori section of recognition classes in the sign space.

The process of constructing models to solve the problem is to choose the most suitable method of machine learning for a particular case with a specific set of data, which is a non-trivial task for which a specialized approach is required. Forecasting the outflow can be considered as a controlled classification problem, in which the behavior of the subscriber is used to teach the binary classifier.

The solution to the problem of customer outflow is reduced to the classification of customers in 2 groups: the customer is prone to outflow or the client is not prone to outflow. If a client falls into the first group, then he must be influenced by the methods of marketing content.

Forecasting the outflow can be considered as a controlled classification problem, in which the behavior of the subscriber is used to study the binary classifier.

The most accurate result is to give the random forest algorithm, because it uses an ensemble of deciduous trees and has the ability to effectively process data with a large number of features and classes.

III. SOURCE DATA

The input data gives us different parameters of the subscribers of the telecommunication network. Parameters have been divide into several groups according to their meaning and purpose of use.

The first group is the main data about the subscriber, which is a general characteristic of the subscriber in the telecommunication network, but do not carry an information component for the analysis (see table 1).

Table I
DATA ABOUT THE SUBSCRIBER, WHICH IS A GENERAL CHARACTERISTIC OF THE SUBSCRIBER IN THE TELECOMMUNICATION NETWORK, BUT DO NOT CARRY AN INFORMATION COMPONENT FOR THE ANALYSIS

Title	Description
ABON_CODE	Hash subscriber code
STATUS	The subscriber's status indicates whether the subscriber is active or not
OBLAST	The area in which the subscriber uses the services of the mobile operator
CITY	The city in which the subscriber uses the services of the mobile operator

The second group provides data about the subscriber's activities in the telecommunication network (see table 2).

One of the complexity of working with data is their different nature, available both numerical and qualitative parameters that are used for the analysis. Additional difficulty is the omission of data. In this case, the correct solution will fill in the missing data with zeros, as this will mean that the subscriber did not use this service or the absence of a card of other operators at the subscriber.

In the analysis, there may be a situation where individual data is redundant and does not affect the output, and even adds an additional error, so these data should not be taken into account during the analysis.

Consideration of the available data parameters allows us to conclude on the different nature of the data and the non-obviousness of their analysis for operators in order to assess their performance.

For each sample of the data for which the analysis should be conduct, it is necessary to determine in detail with all the features present in the input data set in advance.

IV. METRICS

In machine learning tasks to assess the quality of the models and comparison of different algorithms using the following metrics:

– *Accuracy*,

– *Precision*,

– *Recall*,

– Integrated indicator *F-measure*.

Table II
DATA ABOUT THE SUBSCRIBER'S ACTIVITIES IN THE TELECOMMUNICATION NETWORK

Title	Description
CN_OMO_6M	The total number of outgoing calls from other mobile operators
CN_INTL_6M	The total number of outgoing international calls
OP1_DIFF_A_INC_6M	Incoming calls from operator # 1
OR2_DIFF_A_INC_6M	Incoming calls from operator # 2
OP3_DIFF_A_INC_6M	Incoming calls from operator # 3
OTHER_INTL_DIFF_A_INC_6M	Incoming calls from international numbers
DAYS_INACT_ALL_6	Total non-active days
MINS_SLOPE	Tangent of the tilt of the linear trend for the number of minutes weekly
INET_SLOPE	Tangent of the tilt of the linear trend for the number of minutes weekly
INET_STD	The mean square deviation for the data volume weekly
REFILL_STD	The mean square deviation for replenishment weekly
MINS_REG_CONST	”Rise” of the linear trend for the number of minutes weekly
INET_REG_CONST	Line up trend lift for volume of data weekly
REFILL_REG_CONST	”Rise” of the linear trend for replenishment weekly

Accuracy

A dataset is a dimension table m, which consists of parameters where $i = 1, m$. Each i-th parameter in the row p_i of the table takes some values. Thus, each row in the table corresponds to the k-th, where $k = 1, n$, the state of the process, which is analyzed.

In the simplest case, such a metric may be the fraction of states of a set of parameters on which the classifier has made the correct decision.

$$Accuracy = \frac{P}{N}$$

Where, P – number of states of a set of parameters for which the classifier has made the correct decision,

N – the size of the training sample.

In this metric, there is one feature that needs to be taken into account. It consists of assigning all parameters of the same weight, which may be incorrect if the distribution of parameters in the training sample is strongly shift toward one or more classes. In this case, the classifier has more information about these classes and, accordingly, within these classes, it will take more solutions that are adequate. In practice, this leads to the fact that there is an ambiguous definition of metric accuracy for different classes; the discrepancy can range from 80% in a certain class to about 0% in the other.

The solution to this situation is to teach the classifier on a specially prepared, balanced sample of parameters group. The disadvantage of this solution is the loss of information relative to the relative frequency of changes in parameter values.

Precision and *recall*

Precision and *recall* are metrics that primarily use algorithms that require pre-aggregated data as input data. Sometimes these metrics are used separately, and sometimes as a basis for derivative metrics, such as the *F-measure*.

The accuracy of a system within the classroom is the proportion of parameters that really belong to this class among all the parameters that the system attributed to this class. The completeness of the system is the proportion of parameters found by the classifier that really belong to this class in relation to all parameters in the test sample.

The whole set of samples is divid by the classifier into four parts:

- TP (True positive) – the samples are clearly identified by the classifier in a positive class.
- FP (False positive) – the samples are not correctly identified by the classifier in the positive class.
- TN (True negative) – the samples are clearly identified by the classifier in a negative class.
- FN (False negative) – the samples are not correctly identified by the classifier in the negative class.

The dimensions of these parts determine precision and recall:

$$precision = \frac{TP}{TP+FP}$$

$$recall = \frac{TP}{TP+FN}$$

Based on the *precision* and *recall* parameters determine the function used to evaluate the effectiveness of binary classifiers – *F-measure*

F-measure

Of course, the higher the *precision* and *recall*, the better the result. In real life, the maximum values of *precision* and *recall* are not achievable at the same time, so you have to look for a balance. It is advisable to have a universal metric that combines *precision* and *recall* information to evaluate the matching of the algorithm in order to simplify the decision-making process. In this case, the decision process is ask. This metric is the *F-measure*.

The *F-measure* is a harmonic mean between the values of *precision* and *recall*. It should close to zero if the *precision* and *recall* values are approaching zero.

$$F = 2 * \frac{precision*recall}{precision+recall}$$

It is possible to calculate the *F-measure* by giving a different weight to *precision* and *recall* if you determine the priority of one of these metrics during the development of the algorithm.

$$F_\beta = (1+\beta^2) * \frac{precision*recall}{(\beta^2*precision)+recall}$$

Where β takes values in the range $0 < \beta < 1$, when giving priority to *precision*, and when $\beta > 1$, priority is given to *recall*. When $\beta = 1$, the formula is reduced to the previous one and there is a balanced *F- measure*. [9], [10].

The *F-measure* can be used as a formal metric for assessing the quality of a classifier. It reduces two other metrics to one number: *precision* and *recall*. Such a mechanism for assessing the quality of the classifier is much easier to make a decision on the outflow of subscribers.

V. MODEL

For solving classification, problems when using versatile data, the most popular method of machine learning is call *Random Forest*. Random forests are a combination of random trees, so that each tree depends on the values of a random vector taken separately and with the same distribution for all trees of the forest. The generalization error tends to the limit as the number of trees in the forest increases. The error of generalizing the forest of tree classifiers depends on the degree of influence of the generalized indicators of individual trees of the forest and the correlation between them. Internal estimates control the error, the degree of influence of generalized indicators and correlation; they are used to display the answer to an increase in the number of functions used in the splitting. Internal assessments are also use to measure variable importance [11].

Consider the forest construction algorithm.

Stage 1. A subsample of a training sample of a given size is select – a tree is built on it (for each tree there is its own subsample).

Stage 2. To build each splitting (when several edges go out from one node) in the tree, we look through the maximum number of random signs (for each new splitting, its own random signs).

Stage 3. Choose the best sign and carry out splitting according to it (according to a predetermined criterion). A tree is built, as a rule, until the sample is exhausted (until representatives of only one class remain in the leaves), but in modern implementations there are parameters that limit the height of the tree, the number of objects in the leaves, and the number of objects in the subsample at which splitting is performed. This algorithm allows you to determine the signs that have made the greatest contribution, due to the fact that in each node there are known values of the signs.

We will analyze the quality of the solution to the problem of predicting churn subscribers using such mathematical methods:

1) Logistic Regression
2) Bootstrap aggregating
3) Random Forest

We use the Python programming language and the *Pandas*, *Numpy*, *Sklearn*, and *Matplotlib* libraries to build prediction models for subscribers.

Sklearn library contains ready-made algorithms that are use to build models. *Numpy* is use to clear data, structure it and get rid of redundancy. With the help of *Matplotlib*, graphing and visual results are obtaine for further analysis and decision making for each subscriber.

Traditionally solving the problem of customer churn includes:

1) Determination of the circle of subscribers, who are soon ready to abandon the company's services,
2) Establishing the reasons for the refusal of customers from the company's services.
3) Development of cost effective and cost effective measures for their retention.

The results of the prediction regarding the solution of the problem of outflow of subscribers, as well as the analysis of their accuracy and reliability for different algorithms are show in Table 3 and in Fig. 1. An analysis of the results allows us to draw the following conclusions: mid-term methods for finding the best result, showing the method of a hypothesis, results for up to 6% of the total value of the metrics of the most common methods. It surpasses other methods in accuracy, completeness and accuracy of prediction.

For the method of machine learning *Random Forest*, which showed the best results and obtained table 3, showing the influence of the parameters of the telecommunications network on the outflow of subscribers.

In tab. 4 shows the prediction results for the test sample using the *Random Forest* method. The number of correctly and incorrectly estimated data, as well as first and second order errors are show.

Definition of important parameters that affect the outflow of subscribers was performed using the characteristic curve (ROC-analysis), which shows the results of binary classification, when the model assumes the probability that the observation belongs to one of two classes.

Table III
THE METRIC VALUES OBTAINED FOR THE CONSIDERED METHODS OF PREDICTING CUSTOMER CHURN

Model	**precision**	**recall**	**F1**	**F0.5**	**Accuracy**
Logistic Regression	0,709	0,728	0,718	0,713	0,7156
Bootstrap aggregating	0,803	0,75	0,776	0,792	0,7684
Random Forest	0,815	0,7586	0,786	0,804	0,7776

Table IV
SPLIT SET FOR RANDOM FOREST

	importance	labels
Actual False	2046	463
Actual True	649	1842

In tab. 5 shows the 5 parameters that have the greatest impact on the values of metrics characterizing the outflow of subscribers.

Table V
THE IMPORTANCE OF PARAMETERS ON OUTFLOW OF SUBSCRIBERS

Nº	Predicted False	Predicted True
14	0.073648	TENURE
23	0.067708	AVG_DAYS_INACT_ALL_6
21	0.057130	DAYS_INACT_ALL_6
22	0.057023	AVG_DAYS_ACT_ALL_6
30	0.047259	MINS_SLOPE

In fig. 2 shows a histogram of the effect of each parameter on the outflow of subscribers. Based on Fig. 2 and tab. 5, you can determine the parameters that have the greatest impact on the outflow of subscribers. Such parameters are the duration of using the services of a mobile operator, the number of inactive days, the average

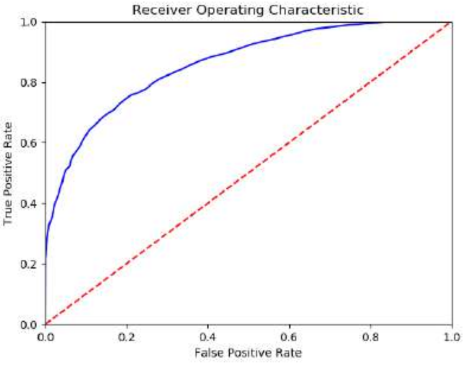

Figure 1. ROC Graph for Random Forest

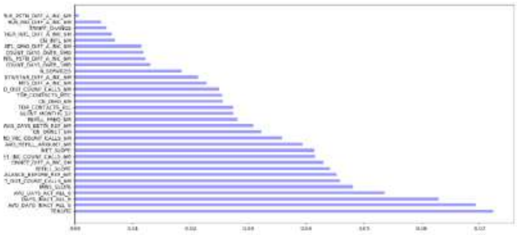

Figure 2. Histogram of parameter dependencies

value of all inactive days, the average value of all active days and the slope of the linear trend for the number of minutes weekly.

Table VI
SUBSCRIBERS PRONE TO CHURN

№	prob_true
13784	1.000000
12763	1.000000
17293	0.914286
10610	1.000000
10555	1.000000

In table. 6 presents data on the likelihood of subscribers to be inclined to outflow and to apply marketing methods to retain them in the network of the mobile operator.

CONCLUSIONS

The paper describes an approach to getting information about the outflow of subscribers of a telecommunications operator. For this purpose, it is propose to use machine-learning methods.

The studies identified a set of subscriber's telecommunications network, which is sufficient for the task. These data form the basis of a process of learning and prediction.

The best results showed a machine learning technique Random Forest which indicators are better to 6-7% as compared to the methods of logistic regression and Bootstrap aggregating.

The result of solving the problem of outflow is:

Identification of the number of subscribers in a short time ready to abandon the services of the company,

Identifying the causes of the failure of customers of the company's services.

Further studies will be focus on increasing the accuracy and completeness of the predictions, as well as on the development of appropriate measures to retain customers.

REFERENCES

[1] Numb: who of mobile operators earned the most. Available at: https://delo.ua/business/onemeli-kto-iz-mobilnyh-operatorov-zarabotal-bolshe-vsego-337433/

[2] Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. SpringerVerlag, 2009. page 746

[3] Arustamov, A. Churn in telecommunication companies. Available at: https://vdocuments.site/9-548b9954b479594c5f8b4658.html

[4] Leo Breiman. Random Forests. Machine Learning October 2001, Volume 45, Issue 1, pages 5–32

[5] Khan, A. A., J. Sanjay, and M. M. Sepehri. 2010. Applying data mining to customer churn prediction in an Internet service provider. Int. J. Comput. Appl. 9(7): pages 8–14.

[6] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach," Eur. J. Oper. Res., vol. 218, no. 1, Apr. 2012, pages 211–229

[7] Y. Sasaki,. The truth of the F-measure. 26th October, 2007 Available at: https://www.toyota-ti.ac.jp/Lab/Denshi/COIN/people/yutaka.sasaki/F-measure-YS-26Oct07.pdf

[8] Telecom Customer Churn Prediction Models. Available at: https://parcusgroup.com/TelecomCustomer-Churn-Prediction-Models

[9] How Ukrainians choose a mobile operator. Available at: https://biz.nv.ua/experts/skorbota/kak-ukraintsy-vybirajut-mobilnogo-operatora-1930612.html

[10] Berkana, A. What is Big data: Collected all the most important information about big data. Available at: https: //rb.ru/howto/chto-takoe-big-data

[11] J. Franklin, "The elements of statistical learning: data mining, inference and prediction," Math. Intell., vol. 27, no. 2, Nov. 2008, pages 83–85

ПРЕДСКАЗАНИЕ ОТТОКА АБОНЕНТОВ ОТ ОПЕРАТОРОВ МОБИЛЬНОЙ СВЯЗИ

Баря А.Д., Глоба Л.С., Мороз А.М.

Аннотация – В данной статье представлен подход к описанию методов машинного обучения для предсказания оттока абонентов оператора связи. Описаны параметры, характеризующие взаимодействие оператора мобильной связи с конечными абонентами. Определены параметры, оказывающие наибольшее влияние на решение клиента об отказе от услуг мобильного оператора. Оригинальность подхода заключается в использовании таких математических методов, которые позволяют определить основной набор параметров, из-за которых конкретные абоненты склонны к смене мобильного оператора. Предложенный подход позволяет организовать систему, в которой возможно определить основные параметры, характеризующие склонность клиентов к оттоку и воздействуя на них разными методами пытаться повысить лояльность абонента. Сравнительный анализ результатов, полученных с помощью проанализированных методов логистической регрессии, беггинга и случайного леса показал, что разброс ошибки предсказания не превышает 6%. Однако преимуществом метода случайного леса является возможность определить набор параметров, которые вносят наибольший вклад в принятие решений абонентом о смене мобильного оператора. Поэтому для проведения анализа, касающегося лояльности абонентов может быть рекомендован метод случайного леса, который показал на тестовой выборке улучшение правильности предсказаний по выборке до 6-7%.

Received 10.12.18