Data Mining Methods for Evaluation and Forecasting the Mobile Internet Traffic in Roaming

Nataliia V. Kuznietsova[a], Petro I. Bidyuk[a] and Anastasiia V. Kulinich[a]

[a] *National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, 03056, Ukraine*

Abstract

This paper is dedicated to the solving such real actual practical task for telecommunication industry as forecasting the services in roaming. It is concentrated on the study of the data mining methods that allow to predict the volume of services (calls, traffic) for a particular subscriber abroad on the bases of available statistical information. The problem was shared in two tasks: forecasting the volume of the internet traffic in roaming and the clients' classification due to their behavior in roaming. The task of evaluation and forecasting was solved with the models based on the time series theory and appropriate autoregression models. The best model was selected based on the statistical criteria and used for forecasting the volume of traffic in next months. The task of classification was solved by such data mining methods as neural networks, gradient boosting, random forest and logistic regression. The model based on the gradient boosting was selected because of the highest completeness and accuracy to the input data. In accordance to received modelling results the recommendations and special strategies for the telecommunication company were developed.

Keywords 1

Data Mining, Gradient Boosting, Random Forest, Neural Networks, Logistic Regression, Bootstrap analysis, Time Series, Mobile Internet Traffic, Forecasting, Roaming

1. Introduction

Today mobile operators are actually interested in resuming their services including mobile roaming, which was observed before the Covid-19 epidemic. In order to predict the mobile Internet volume, including abroad [1], it is necessary to determine optimal packages and behavior of subscribers [2, 3], i.e. how they use certain services within their own country. In the future, it will be possible to predict how such subscriber will behave abroad using modern methods of data mining and forecasting [4 – 12].

Special mention should be paid to virtual mobile operators (for example, LycaMobile), which use roaming technology and rent telecommunication towers and equipment of other mobile operators to provide services. Virtually every subscriber of a virtual operator is in roaming within his own country, and therefore information on the Ukrainian services usage can be used to predict his behavior abroad.

Roaming is a situation when a mobile operator subscriber (of the home network) uses the network of another mobile operator outside the geographical coverage area of the home network. The peculiarity of doing business in the field of international roaming is the need to make a huge number of deals (agreements) with major international telecommunication operators providing services in more than a hundred countries [1]. The national operator should take into account that certain services in some telecommunication companies abroad are extremely expensive for both the operator and the subscriber. Thus, roaming tariffs in such countries should be formed and allocated

CMIS-2021: The Fourth International Workshop on Computer Modeling and Intelligent Systems, April 27, 2021, Zaporizhzia, Ukraine

EMAIL: natalia-kpi@ukr.net (N. Kuznietsova); pbidyuke_00@ukr.net (P. Bidyuk); a.kulinich2509@gmail.com (A. Kulinich)
ORCID: 0000-0002-1662-1974 (N. Kuznietsova); 0000-0002-7421-3565 (P. Bidyuk); 0000-0001-7156-109X (A. Kulinich)

in a separate group, because for the mobile operator one minute can cost much more than the company can receive from the subscriber.

The national mobile operator must also provide appropriate conditions for the usage of mobile services for its subscribers around the world. Services that subscribers use abroad can be divided into several types including voice communication, messages and mobile Internet.

Usually Ukrainian subscribers abroad prefer voice calls and mobile data usage. It was expected that with providing new 4G standard the traditional telecommunication services such as voice calls and short messages (SMS) remained in the past. Now there is an actual decrease in the number of subscribers who use standard short messages but with voice calls it is not so clear. The people needs in live communication which is impossible in today’s lockdown and quarantine situations have led to a huge increase in demand for voice calls. Therefore, now this service remains one of the most developed and necessary for users and therefore relevant in roaming. Let's take a closer look at the tasks of forecasting services and tariff packages faced by telecommunication companies to provide users in roaming. To do this, first the problem of the traffic in roaming amount forecasting will be solved. Next the subscribers’ classification according to mobile Internet tariff packages in roaming will be performed. As a result, it will be possible to determine which tariff packages are relevant now for subscribers, whether it is needed to develop one or more different proposals.

2. Problem statement

This work is concentrated on the study and research of the information tools that allow one to predict the volume of services (calls, traffic) for a particular subscriber abroad on the bases of available statistical information. Mathematical models need to be developed to analyze consumption and forecast the services that will be used abroad. It is necessary to predict the mobile Internet traffic volume in roaming using the statistics on subscriber behavior and perform the subscribers’ classification in order to develop relevant packages and offers. Finally it is needed to develop special techniques and approaches to increase the number of subscribers who will use these services in roaming.

3. The main methods used in the research

According to the seventh international workshop held by SAS in September 2020 "Real Time Analytics & Cyber Security 2020", the "big four" of the most promising and relevant methods for solving analytical problems in financial sector were as follows: neural networks [11– 13], gradient boosting, random forest and support vectors machine [14 – 16]. It was also noted that the high results for classification shows the logistic regression method. So, based on the international experience we chose these methods to solve the problem of classification for our tasks. For the internet traffic forecasting the theory of time series analysis [8, 9, 17, 18] was applied and some autoregressive models were explored and built after preliminary data processing.

3.1. Logistic regression

Logistic regression is a statistical method used to analyze a data set consisting of one or many independent characteristics that affect the outcome. The result is evaluated using a dichotomous variable, which indicates with what probability the result belongs to a particular class [10]. The logistic regression algorithm uses a linear equation (boundary function) with independent variables to determine which class the data belongs to [4]. This equation describes the linear boundary that separates the input data space.

The limit function can be generally written as follows:

$$Score(x_i) = w_0 + w_1 x_1 + ... + w_n x_n = w^T \cdot x_i, \tag{1}$$

where x_i is input variable feature, w_0 is decision making threshold, $w_1,...,w_n$ is vector of weights.

In order to obtain the probability value [0; 1] of the class membership from the boundary function, the following logistic function is used:

$$\log it(Score) = \frac{1}{1-e^{-Score}}. \tag{2}$$

First of all, to construct the limit function it is necessary to find the coefficients, $w_1,...,w_n$. To do this, it is necessary to determine the training sample, which consists of independent variables (characteristics) and the corresponding values of the dependent variable y (initial result). Formally, it is a set of pairs, $(x^{(1)};y^{(1)})...(x^{(m)};y^{(m)})$, where $x^{(i)} \in R^n$ is the vector of independent variables values and $y^{(i)} \in \{0,1\}$ is the corresponding value of y. Each such pair is called the learning example. Usually the method of maximum likelihood is used and the parameters are selected so that to maximize the value of the likelihood function in the training sample [4]:

$$W = \arg\max_w L(W) = \arg\max_w \prod_{i=1}^{m} P\{y = y^{(i)} \mid x = x^{(i)}\}. \tag{3}$$

Maximizing the likelihood function is equivalent to maximizing its logarithm:

$$\begin{aligned} \log L(W) &= \sum_{i=1}^{m} \log P\{y = y^{(i)} \mid x = x^{(i)}\} = \\ &= \sum_{i=1}^{m} y^{(i)} \log f(w^T x^{(i)}) + (1-y^{(i)})\log(1-f(w^T \cdot x^{(i)})) \end{aligned} \tag{4}$$

where $x^{(i)} = w_0 + w_1x_1 + ... + w_nx_n$.

To maximize this function the gradient descent method could be used. By setting some initial value w_0 a maximum can be found iteratively [4]:

$$w = w_0 + \alpha\nabla \log L(w) = w_0 + \alpha\sum_{i=1}^{m}(y^{(i)} - f(w^T x^{(i)}))x^{(i)}, \alpha > 0. \tag{5}$$

3.2. Neural networks

The most modern neural networks are constructed of formal neurons that resemble their biological prototype. The structure of the neuron consists of $x_1,...,x_n$ are the values that are fed to the inputs (synapses) of the neuron; $w_1,...,w_n$ are weighting coefficients of synapses, which can have both slowing down and strengthening effect; S is the weighted sum of the input characteristics:

$$S = \sum_{i=1}^{n} w_i \cdot x_i - T, \tag{6}$$

T is neuron threshold (omitted in many models), F is the neuron activation function that converts the weighted sum into an output signal: $y = F(S)$ [20].

The neurons are regularly organized into layers, and the elements of a layer are associated only with the neurons of the previous layer and the information spreads from the previous layers to the next. The input layer, which consists of sensitive (sensory) S-elements, which receives the input signals, X_i, does not perform any information processing and performs only distribution functions. Each S-element is associated with a set of associative elements (A-elements) of the first intermediate layer, and the A-elements of the last layer are connected to the reacting elements (R-elements) [20].

Weighted combinations of R-element outputs determine the system response which indicates that the evaluated object belongs to a certain image. If only two images are recognized, then one R-element is installed in the perceptron, which has two reactions – positive and negative. If there are more than two images then for each image its R-element is set and each such element output is a linear combination of output A-elements.

Neural networks are now one of the most common methods that is being developed, and adapted, using recurrent algorithms (RNN), long short-term dependency learning (LSTM), to solve regression problems (GRNN). Therefore, the use of neural networks is appropriate for our problems.

3.3. Random Forest

The random forest method is based on the large number (ensemble) of decision trees (this number is a parameter of the method) construction, each of which is built on a sample obtained from the original training sample using bootstrap (i.e. sample with return), in contrast to classical construction algorithms decision trees [15].

The *bootstrap procedure* is randomly retrieving repeated samples from the empirical distribution multiple times. Specifically, if we have an initial sample of n terms, $x_1,.x_2,..,x_{n-1},x_n$, then by using the random numbers generator evenly distributed on the interval $[1,n]$, we can extract from it an arbitrary element x_k, which will be returned to the original sample for possible re-extraction. This procedure is repeated n times. A bootstrap sample is formed where some elements can be repeated two or more times, while other elements are absent. For example, for $n = 6$ one of such bootstrap combinations has the following form: x_1,x_2,x_2,x_1,x_4,x_5 [21].

Bootstrap samples are performed evenly and with a return, so some initial samples will be missing while others will be duplicated: on average, one such sample contains about 2/3 of unique initial observations. Bootstrap was particularly useful in models ensemble formation especially in combination with tree-like structures which are very sensitive to small changes in training data [6].

As the averaging of several observations reduces the data variance estimation the same reasonable way to reduce the variance of the forecast is to obtain a large number of data from the general population, building a predictable model for each training sample and averaging of the obtained forecasts. If instead of separate training samples to perform bootstrap and based of the generated pseudo-samples to build B regression trees, the average collective forecast will have a lower variance:

$$\hat{f}_{bag} = (f^1(x) + f^2(x) + ... + f^B(x))/B. \qquad (7)$$

This procedure is called *bagging* (short abbreviation for bootstrap aggregating). Bagging can be performed not only for regression trees but also to other models [10].

The random Forest is an improvement of decision tree bagging which aims to eliminate the correlation between trees. As with bagging we build several hundred decision trees based on training bootstrap samples. However, at each iteration of constructing the trees are randomly selected m from p as to be considered as predictors and it is allowed to perform partitioning only by one of these m variables [15]. The meaning of this procedure is quite effective for improving the quality of the obtained solutions and it is that with the probability $\frac{p-m}{p}$ any potentially dominant predictor that seeks to enter each tree is blocked. If the dominance of such predictors is allowed then all the trees as a result will be very similar to each other. Also the obtained on their basis forecasts will be strongly correlated and the decrease in variance will not be so obvious. By blocking the dominants other predictors will get their chance and the tree variation increases. Choosing a small value of m when constructing a random forest will be useful in case of a large number of correlating predictors. Naturally, if a random forest is built using $m = p$ then the whole procedure is reduced to a simple bagging [20].

Random forests provide a significant increase in accuracy while the trees in the ensemble are weakly correlated due to the double injection of randomness into the inductive algorithm – by bagging and random subspaces methods for splitting each vertex; they don't exhibit the overfitting problem. They are easy for usage: the only algorithm parameters are the trees number in the ensemble and the number of traits randomly selected for splitting at each top of the tree.

3.4. Gradient Boosting

Now let's consider the problem of recognizing objects from the multidimensional space X with the label space. Let a training sample, $\{x_i\}_{i=1}^N$, where $x_i \in X$ is given. And let are known the true values of the labels for each object, $\{y_i\}_{i=1}^{N}$, where $y_i \in Y$. It is necessary to build a recognition operator that can predict the labels for each new object $x \in X$ as accurately as possible. Let the family of the basic algorithms H is given, each element of $h(x;a) \in H : X \to R$ is determined by some vector of parameters $a \in A$ [7].

We will search for the final classification algorithm in the form of the composition

$$F_M(x) = \sum_{m=1}^{M} b_m h(x;a_m), b_m \in R,\ a_m \in A. \tag{8}$$

However, the selection of the optimal set of parameters $\{a_m, b_m\}_{m=1}^{M}$ is a very time-consuming task. Therefore, we will try to build such a composition by greedy manner building, each time adding to the sum a term which is the most optimal parameter from all possible. We assume that we have already constructed a classifier F_{m-1} of length, $m-1$. Thus, the problem is to find a pair of the most optimal parameters $\{a_m, b_m\}$ for the classifier of length, m:

$$F_m(x) = F_{m-1}(x) + b_m h(x;a_m), b_m \in R, a_m \in A. \tag{9}$$

The idea of boosting can be also applied to classification. In case of binary classification, this means $Y = \{-1;+1\}$. Then it is often assumed that each algorithm $h \in H$ returns the actual «degree» of object belonging to a certain class, and the resulting answer $\widetilde{F}$ is obtained by applying a boundary rule to the composition [16].

3.4.1. Multiclass classification

The idea of boosting for binary classification could be easily generalized to the case of K classes [22]. Now the following loss function is introduced:

$$L(y,F) = -\sum_{i=1}^{k} y_i \log p_i(x). \tag{10}$$

Here $y_i \in \{0,1\}$ shows the affiliation of the object of class i, and p_i shows the probability of belonging the object to the class i, obtained during application of the logistic regression. Write down the formulas for the class K classifier of multiclass logistic regression:

$$f_k(x) = \log p_k(x) - \frac{1}{K}\sum_{i}^{K} \log p_i(x). \tag{11}$$

It is possible to receive after transformations that:

$$\nabla Q_i = y_{ik} - p_{k,m-1}(x_i). \tag{12}$$

If the problem is too complicated for calculations then in case of computational trees it is possible to use the first step of the Newton-Raffson algorithm as an approximation:

$$c_{jm} = \frac{K-1}{K} \frac{\sum_{x_{j \in R_{jkm}}} \nabla Q_{ik}}{\sum_{x_j \in R_{jkm}} |\nabla Q_{ik}|(1-|\nabla Q_{ik}|)}. \tag{13}$$

The i-th classifier search for such object class means that the probability of belonging of other classes was minimum:

$$c_{jm} = \underset{k \in [1,K]}{\arg\min} \sum_{k=1}^{K} c(k,\widetilde{k}) p_{\widetilde{k}m}(x). \tag{14}$$

In this formula, $p_{\widetilde{k}m}(x)$ denotes the probability of belonging to class k as a result of the m-th iteration of the boosting algorithm. The value of $c(k,\widetilde{k})$ denotes the error cost function if it is assumed that the object belongs to the class, k, although in fact it belongs to the class $\widetilde{k}$ [16].

4. Input data features and characteristics description

To analyze and predict subscribers who are using roaming services, the mobile operator provided a sample of 120,000 records – data for subscribers traveling abroad. For 10,000 random records from each month for the period from August 2017 to July 2018, i.e. for twelve consecutive months. It is necessary to predict, first of all, whether the subscriber traveling abroad will use communication services as well as to determine which services (calls or mobile internet) the subscriber will use.

The input data contains the following characteristics: the previous month before going abroad (used to display the history of using services in Ukraine); and a set of characteristics that reflect the features and types of services used by the subscriber in Ukraine (subscriber's internal tariff plan, the number of calls minutes, the number of short messages, the amount of GPRS traffic, the cost spent by the subscriber to use services in Ukraine, the amount and quantity replenishment of the balance in the specified month, administrative region of Ukraine, in which the subscriber was more than 90 days before the moment of departure abroad). We also add characteristics for our task: the month in which the subscriber was left abroad, the code of the country tariff group and the name of the country to which the subscriber left.

The following characteristics describe the amount of services that the customer used in roaming during his previous trip abroad. Such information is stored by the mobile operator for the entire period of customer service, but for our task it is advisable to take into account the previous visit not earlier than the year of 2016. This is due to significant changes in tariffs and the high cost of roaming services, and therefore there was very infrequent use of this service before 2016. This is significantly different from today's situation and the mobile operator faces the task of determining the subscriber behavior and relevant services in modern conditions.

- HISTORY_ROAM_GPRS is the amount of GPRS traffic in roaming.
- HISTORY_ROAM_MINS is the amount of call minutes in roaming.
- HISTORY_ROAM_SMS is the number of short messages in roaming.

If the subscriber has not traveled abroad in the last three years, the characteristic values will be empty. Empty values of these characteristics will be further processed and filled by one of the methods for incomplete data recovering (average or median value for the tariff group of countries) [23].

In total the sample contains 15 characteristics that describe the subscriber behavior in Ukraine and abroad. An experimental study was conducted and such characteristics from behavioral characteristics in Ukraine were generalized and selected for further modeling:

- MO_UKR is the sum of outgoing calls made inside the network to other operators and also outgoing calls abroad.
- SMS_UKR is the sum of outgoing short messages in the network sent on the phone numbers of the national operators and international short messages.
- AMOUNT is the total sum of all services interaction into calls, GPRS and short messages.

5. Data mining methods application to solve the classification and forecasting problems

5.1. The task of predicting the services amount that will be used by the subscriber in roaming

Based on subscribers' statistics it is necessary to forecast and offer the outgoing subscriber an appropriate services package for roaming. It is also necessary to predict the amount of mobile GPRS traffic in roaming, which will be used by the subscriber – the target variable Y_mi. It is advisable to try to solve this problem by using different types of regression models [5, 8, 18].

To forecast the mobile data in roaming usage was used the dataset for the period from January 2016 to July 2018. Altogether, the data was provided for 31 consecutive months. A graphical representation of the time series for mobile internet usage in this period is shown in Figure 1.

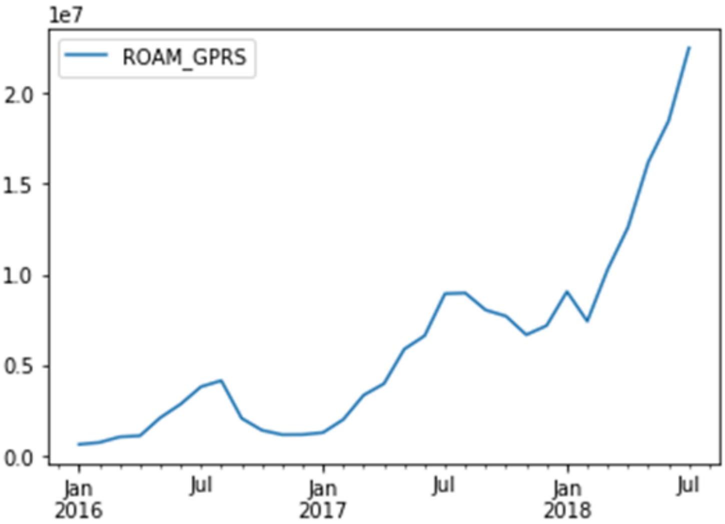

Figure 1: The volume of GPRS mobile data traffic, MB

The usage of the GPRS service abroad is growing all this period due to a significant reduction in the prices for the service of subscribers and international partners. Visually, the time series shown in Figure 1, has an exponential trend, but we still firstly conducted a series of studies on stationarity, constructed a moving average and standard deviation, performed Dickey-Fuller and KPSS tests, which indicated that the time series is not stationary (Dickey-Fuller: p-value = 1.00, KPSS criterion: p-value = 0.05) [24, 25].

In order to transform the original series of mobile data in roaming to stationary, it is necessary to remove the exponent by usage the second differences. After this transform the time series became stationary so autoregression models and autoregression with moving average (MA) models could be used. For building this models we need to find the order of the model and the order for moving average.

The process of parameter selection for the models $AR(p)$, $MA(q)$ and $ARMA(p,q)$ for the series is based on partial autocorelation functions to determine the lag and search for the parameters p and q. The results of comparative analysis of estimating statistical characteristics for the selected models are presented in table 1.

Table 1
Comparison of the models which describe the process of using mobile data traffic in roaming

Models	Criteria			
	R^2	AIC	BSC	DW
AR(2)	0.95	843	847	1.97
MA (6)	0.95	699	709	1.76
ARMA (2, 6)	0.98	676	689	2.12

The best model was defined the model ARMA (2, 6). In Figure 2 are shown the results of simulation based on ARMA (2, 6) with the input data values for a given series. Next based on this model the forecast for the next 7 months ahead was built and provided to the mobile operator for estimation of further tariff strategy.

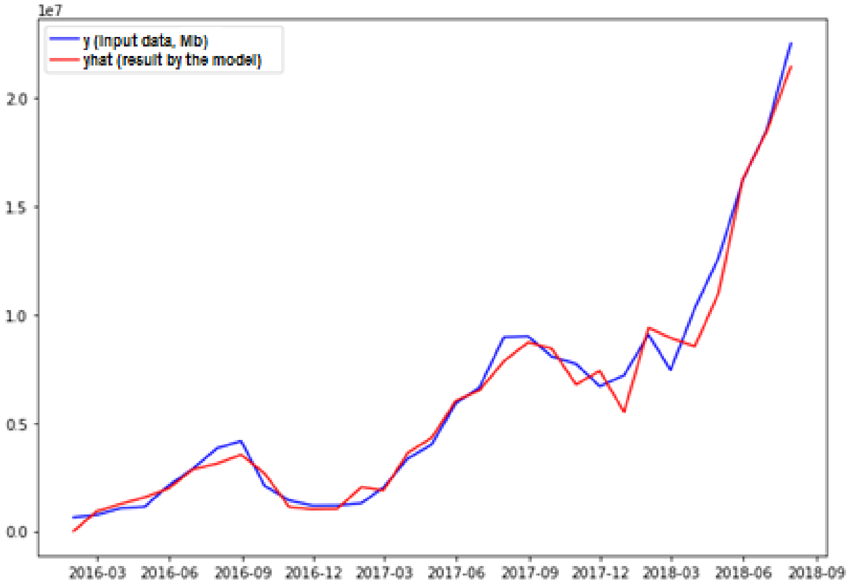

Figure 2: The results of the ARMA model (2, 6) in comparison with the original series

5.2. Classification models construction to determine the best package of services for roaming subscribers

To develop new tariffs and service packages, a sample of 120,000 records which contained subscriber's data who were staying abroad in the period from August 2017 to July 2018 was provided. The task was to classify the target variable – a package of services for mobile data transmission in roaming by classification models.

Let's preliminary define the main service packages for GPRS mobile internet in roaming. The target variable Y_p_mi for predicting GPRS traffic in roaming can take the following values:

- gr_0 – the subscriber did not use the service at all;
- gr_100 – the subscriber used up to 100 MB;
- gr_500 – the subscriber used from 100 to 500 MB;
- gr_over_500 – the subscriber used more than 500 MB.

It is known that subscribers when traveling abroad often choose not to use mobile services at all. For calls in roaming such group of subscribers is near 66%, for GPRS services – 42% (in 50,623 trips subscribers will not use any of the internet service packages). Figure 3 shows the distribution of classes for different mobile Internet services in roaming packages.

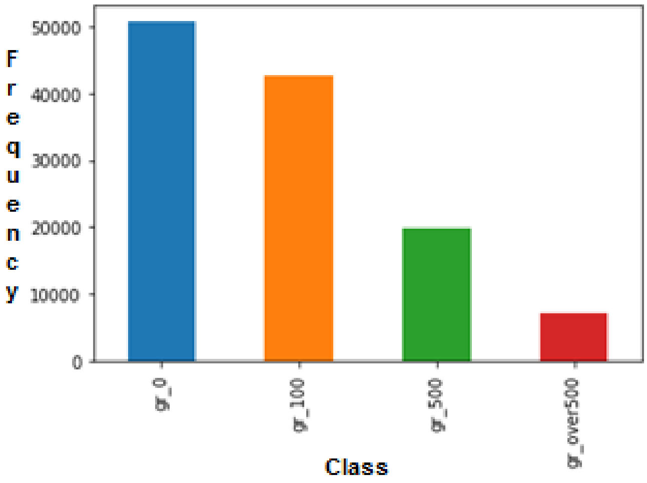

Figure 3:The frequency of the mobile data transmission in roaming usage classes

The data on the above mentioned 15 characteristics were formed by using the appropriate technical systems of the mobile operator, the numerical values are accurate, and the term data has a limited finite number of valuesand can be immediately taken as the categorical features.

5.2.1. Data processing and preparation

For continuous variables in the input dataset it was used discretization – splitting a continuous variable into some categories [25]. To convert the categorical features represented by string literals the direct coding or as it is also called one-hot-encoding (coding with one active state) was used [4]. The basic idea is to replace a category variable with a new variable by the formula for a linear binary classification. In such a way N new features are created for a categorical feature, where N is the number of categories. Each new feature is a binary characteristic of a certain category.

After preliminary data processing initial sample was divided into training (75%) and test (25%) data set. While the sample is unbalanced the class stratification method was used to ensure that sufficient values were obtained from each group.

On the next step the sample was balanced by increasing the number of records of smaller classes to the number of the majority class, increasing their weight for the training sample. A modified method of over-sampling the SMOTE (Synthetic Minority Over-sampling Technique) [19] was used, in which records are not simply duplicated but artificially generated based on examples of real representatives of the class with minor deviations.

5.2.2. The statistical characteristics for checking the classification quality used in our task

The most well-known statistical indicators for assessing the quality of classification are general accuracy, error matrix (Confusion Matrix), first and second kind of errors, index GINI [25]. Usually the error matrix is determined for binary classification, size 2×2. In general, the error matrix $N \times N$, where N is the number of classes, which shows the correct model predictions, as well as forecast errors. For our classification problem into four classes the matrix is presented in the form of the following table 2. Here the classification errors are divided into the following groups:

FN (False Negative) – first kind of errors, false negative value. For our task such error means that the forecast value implies a package of services smaller than actually needed.

FP (False Positive) – second kind of errors, false positive value. Here, this error means that the forecast value implies a package of services greater than actually needed [22].

Table 2
Confusion Matrix for mobile internet in roaming

		Forecasted values			
		$\hat{y}=1$	$\hat{y}=2$	$\hat{y}=3$	$\hat{y}=4$
Real values	$y=1$	TP	FP	FP	FP
	$y=2$	FN	TP	FP	FP
	$y=3$	FN	FN	TP	FP
	$y=4$	FN	FN	FN	TP

The peculiarity of the classification task of the package services is that the fact that the second kind errors are more acceptable for marketing decisions than errors of the first kind. Because the first kind of errors will mean potentially unearned income, and mistakes of the second kind – only a failed attempt to sell a package of services.

The values of the confusion matrix are used to calculate the main metrics for estimating the classification model.

The overall *accuracy* of the model is calculated as [26]:

$$accuracy = \frac{TP}{TP + FP + FN}. \tag{15}$$

To evaluate a model with unequal classes this metric will not fully characterize the model correctness. On example of the subscribers' behavior in roaming classification it means that if 70% of subscribers do not use roaming services, then even if all instances are classified as class 0, so the accuracy of the model equal to 0.7 will be obtained. Obviously, the purpose of the forecast does not coincide with this result and the model doesn't have significant practical value.

The quality assessment of the multiclass classification model is performed using the following metrics, which are calculated for each class separately.

Precision shows that some of the objects which were called positive by the classifier and in fact they were indeed positive.

$$precision = \frac{TP}{TP + FP}. \tag{16}$$

Recall shows which part of the positive class objects out of all the positive class objects were found by the algorithm.

$$recall = \frac{TP}{TP + FN}. \tag{17}$$

The recall demonstrates the ability of the algorithm to find this class in general, and precision – the ability to distinguish this class from other classes [26].

F-measure (f1-score) is the average harmonic between precision and recall:

$$f_\beta = (1+\beta^2)\cdot\frac{presicion \cdot recall}{(\beta^2 \cdot precision) + recall}, \tag{18}$$

where β determines the accuracy measure in this metric and for $\beta = 1$ it is the harmonic mean (with the factor $(1+\beta^2)$ equal to 2, so that in the case when $precision = 1$ and $recall = 1$) the f1-score reaches its maximum at completeness and accuracy equal to one. The measure f1-score approaches zero if one of the indicators approaches zero [26].

The *Matthews correlation coefficient* (MCC) in contrast to the previously considered estimates takes into account all the values of the confusion matrix and is calculated by the following equation:

$$. \qquad MCC = \frac{c \cdot s - \sum_{k=1}^{K} p_k t_k}{\sqrt{(s^2 - \sum_{k=1}^{K} p_k^2)(s^2 - \sum_{k=1}^{K} t_k^2)}}, \tag{19}$$

where t_k is the number of instances in class k; p_k – how many times the model predicted the class k; c – is the number of correctly classified instances; s – is the total number of instances.

The MCC can take the values from -1 to +1. A model rated to +1 is considered as an ideal. The model which received a score of -1 is considered very weak.

6. Simulation results

The simulation was performed on the basis of such methods as: logistic regression, neural networks, random forest, and gradient boosting. Let's consider more in detail the results of application of the methods described above to classify the proposals of the mobile data service in roaming. Visually the results of each model in the test sample are shown by using the confusion matrix in Figure 4.

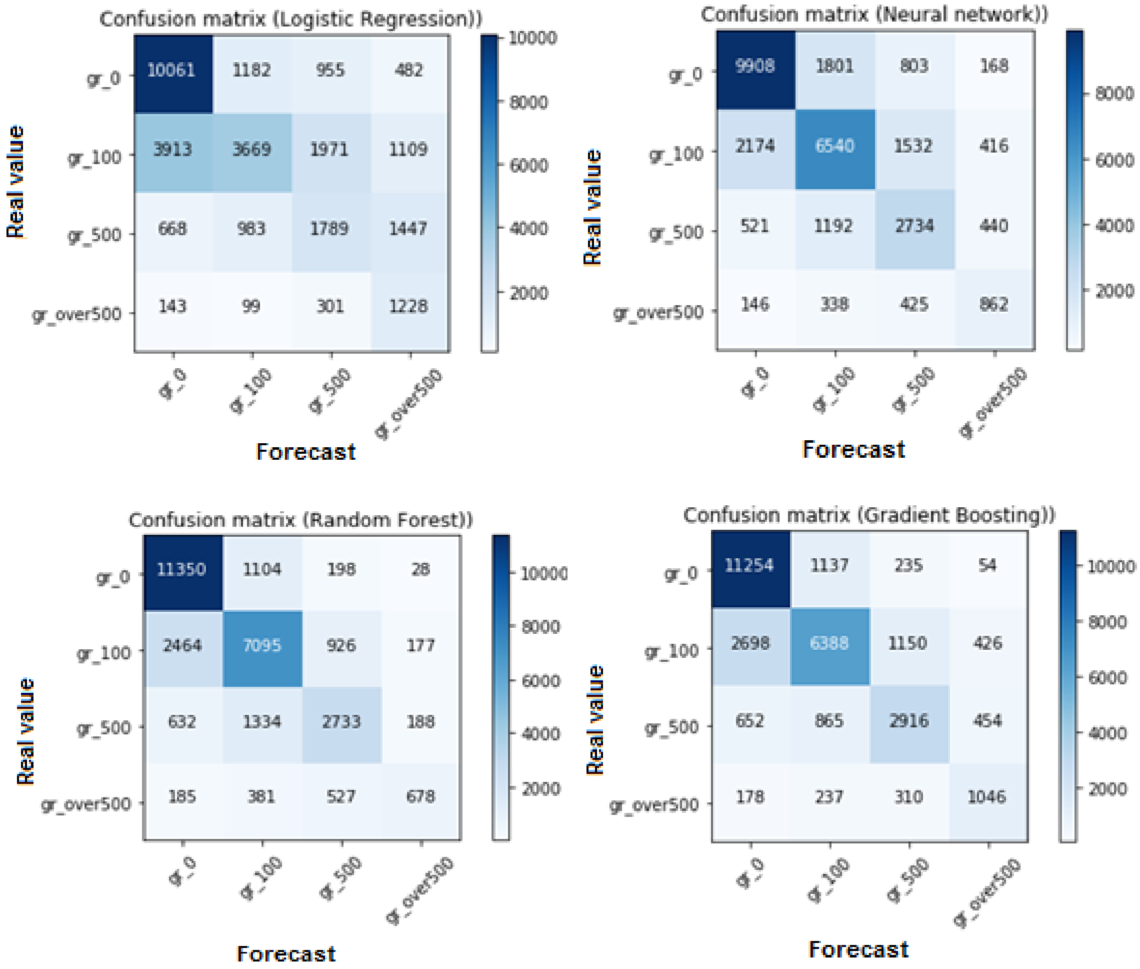

Figure 4: Confusion matrixes for the mentioned above models for mobile data services in roaming classification

GPRS service is gaining more popularity among the national mobile operator subscribers not only in Ukraine but also in roaming. In the case of offering such services one should expect a better response on this offers from subscribers even if they are from the erroneously predicted model classes (FP - errors). To select the best model we will use similar indicators of the classification models quality. The results for all models are gathered together in Table 3.

Final comparisons of the classifiers were made using the Matthews coefficient: logistic regression shows MCC=0.36; Neural Network MCC=0.51; Random forest – MCC=0.58; and Gradient Boosting – MCC = 0.58. The results of the Gradient Boosting and Random forest models are very similar in all respects. Only f1-score for Gradient Boosting has a few percent better performance, so this model can be considered as the best one.

7. Discussion

According to the simulation results received, several possible improvements in the formation of proposals for subscribers traveling abroad were proposed.

The first scenario takes into account the current state of the roaming communications market and aims to analyze customer preferences and offer appropriate service packages. Such proposals may provide some extra benefits for clients. For example, such as:

Table 3
Comparison of classification models for the use of GPRS in roaming

Class	Method	precision	recall	f1-score
gr_0	Logistic regression	0.68	0.79	0.73
	Neural Network	0.78	0.78	0.78
	Random forest	0.78	0.90	0.83
	Gradient Boosting	0.76	0.89	0.82
gr_100	Logistic regression	0.62	0.34	0.44
	Neural Network	0.66	0.61	0.64
	Random forest	0.72	0.67	0.69
	Gradient Boosting	0.74	0.60	0.66
gr_500	Logistic regression	0.36	0.37	0.36
	Neural Network	0.50	0.56	0.53
	Random forest	0.62	0.56	0.59
	Gradient Boosting	0.63	0.60	0.61
gr_over 500	Logistic regression	0.29	0.69	0.41
	Neural Network	0.46	0.49	0.47
	Random forest	0.63	0.38	0.48
	Gradient Boosting	0.53	0.59	0.56

- for subscribers whose behavior is classified in the group gr_100 to offer 150 megabytes in the second purchased package. This encourages the subscriber to use more than 100 megabytes and to order one more package;
- for subscribers who are classified by the model in the group gr_500 (the subscriber used from 100 to 500 megabytes) to offer to purchase till promotional date a package of 500 megabytes with a 5% discount;
- for subscribers who have been classified by the model in the gr_over500 group to offer a 1 GB service package instead.

This will encourage subscribers who use roaming services to use a little more services, and thus improve the experience and encourage the quantity of people who use the roaming services and thus to increase the total amount of mobile traffic in roaming.

The first scenario can be implemented using classification models. The best model based on the Gradient Boosting provides satisfactory accuracy and completeness of classification, which gives confidence that the target audience for the marketing campaign will be defined qualitatively.

The second scenario involves a significant marketing campaign aimed at reducing the percentage of subscribers who do not use mobile services when traveling abroad. The first assumption: subscribers cannot assess their own need to use services, estimate costs and therefore prefer not to use services at all, sometimes even turn off their phones.

A successful offer for a fixed package of services, for example, 100 megabytes with the specified cost and a guarantee that exceeding this limit is impossible and will not incur excessive costs can positively affect the decisions of subscribers. The main task is to determine among subscribers who do not use roaming services, their behavior characteristics in Ukraine which will give us assumption which package should be offered and to whom. For this purpose were developed and trained the models mainly on the subscribers' behavior who use roaming services.

For the second scenario, one can also use the proposals similar to the first scenario. The equality of the second scenario is that you need to understand that the cost of a marketing campaign can be much higher and the conversion much lower. The second scenario is designed for the long term campaign.

All developed models help to qualitatively form and determine the target audience for the marketing proposals. The decision to provide special offers or inform subscribers is made by marketing department depending on the available budget.

8. Conclusions

Mobile operators perform behavioral users' analysis in order to develop new tariff packages and retain existing subscribers among their active users. The Covid-19 crisis has shown that the modern world is adapting to new conditions at an extremely rapid pace and there is a need to support and develop telecommunications and Internet services. After the borders are opening between countries and intensification of the international travel which may take place in early 2022 the tourist and business flow will resume to the pre-crisis level. It is expected that during the year it will reach the indicators of 2018 year, so the simulation and forecasting of roaming traffics will allow mobile operators to develop appropriate tariff policies for roaming subscribers.

From the all four built classification models for predicting the subscribers behavior in roaming a model based on the gradient boosting was selected because of the highest completeness and accuracy to our input data. This model should be used for prudent marketing strategies. For bolder marketing campaigns such classification model which will contain more second kind errors should be used. In such case it is predicted that the subscriber should be offered a tariff but in fact he will not use the service. Such mistakes will allow to conduct a marketing campaign to activate roaming subscribers who do not use the services. In conditions of increasing numbers of subscribers that get used Internet the chances of success for such marketing campaign will increase day by day.

In further research it is planned to analyze and forecast the usage by clients other services in roaming (calls and SMS), as well as in cooperation with other departments of the mobile operator to give recommendations on the company's development and distribution of various services in national and international vectors.

9. References

[1] European Commission Press Release. Brussels, 17 February 2014. URL: http://europa.eu/rapid/press-release_IP-14-152_en.htm.

[2] N. V. Kuznietsova, Information Technologies for Clients' Database Analysis and Behaviour Forecasting, in: CEUR Workshop Proceeding, 2017, pp. 56-62. URL: http://ceur-ws.org/Vol-2067/.

[3] M. Havrylovych, N. Kuznietsova, Survival analysis methods for churn prevention in telecommunications industry, in: CEUR Workshop Proceeding, 2020, pp. 47-58. URL: http://ceur-ws.org/Vol-2577/paper5.pdf.

[4] G. James, D. Witten, T. Hastie, R. Tibshinari, An Introduction to Statistical Learning with Applications in R, Springer-Verlag, New York (2013). doi: 10.1007/978-1-4614-7138-7.

[5] N. S. Papageorgiou, V. D. Radulescu, D. D. Repovs, Noninear Analysis – Theory and Methods, Springer, Cham, Switzerland, 2019.

[6] L. Breiman, Random forests, Machine Learning 45 (2001), 5–32. URL: https://doi.org/10.1023/A:1010933404324.

[7] J. Friedman, Greedy Function Approximation: A Gradient Boosting Machine, The Annals of Statistics, vol.29, No.5 (2001), pp. 1189-1232. URL: https://www.jstor.org/stable/2699986?origin=JSTOR-pdf&seq=1 .

[8] T. M. Rassias, Applications of Nonlinear Analysis, Springer, Cham Switzerland, 2018.

[9] J. Beran, Mathematical Foundations of Time Series Analysis, Springer, Cham Switzerland, 2017.

[10] V. K. Shitikov, S. E. Mastitsky, Classification, regression and other data mining algorithms using R(in Russian), 2017. URL: https://ranalytics.github.io/data-mining/index.html.

[11] N. V. Kuznietsova, M. Seebauer, S. Zabielin, Some methods for estimating financial risks in banking, IEEE 1st Conf. on System Analysis and Intelligent Computing, SAIC 2018 (2018), pp. 271–274. doi: https://ieeexplore.ieee.org/document/8516873.

[12] S. Osovsky, Neural networks for information processing (translation in Russian by I.D. Rudinsky), Finance and statistics, Moskow, 2002.

[13] O.V. Gorokhovatsky, O. O. Peredriy, Multilayer perceptron as the primary instrument for image clustering (in Ukrainian) / Registration, storage and data processing 18 (2016) 33–43.

[14] C. Cortes, V. Vapnik, Support-vector networks, Machine learning (1995), vol. 20, no. 3, 273–297.

[15] S.P. Chistyakov, Random forests: an overview (in Russian), Works of Karelian scientistist center of RAS, 2013, Issue. 1, pp. 117–136. URL: http://resources.krc.karelia.ru/transactions/doc/trudy2013/trudy_2013_1_117-136.pdf.

[16] P. Tseng, On accelerated proximal gradient methods for convex-concave optimization, 2008. URL: http://www.math.washington.edu/□tseng/papers/apgm.pdf.

[17] R. S. Tsay, Analysis of financial time series, John Wiley & Sons, Inc., New York, NY, 2010.

[18] P. Bidyuk, A. Gozhyj, Y. Matsuki, N. Kuznetsova, I. Kalinina, Modeling and Forecasting Economic and Financial Processes Using Combined Adaptive Models, in: Babichev S., Lytvynenko V., Wójcik W., Vyshemyrskaya S. (Eds.), Lecture Notes in Computational Intelligence and Decision Making, ISDMCI 2020, Advances in Intelligent Systems and Computing, vol 1246, Springer, Cham, 2021. URL: https://doi.org/10.1007/978-3-030-54215-3_25.

[19] V. N. Nikulin, I. S. Kanishchev, I.V. Bagaev, Methods of balancing and normalization of data to improve the quality of classification, Computer tools in education 3 (2016) 16–24.

[20] V. K. Shitikov, G. S. Rosenberg, T.D. Zinchenko, Quantitative hydroecology: methods of systemic identification, IEVB RAS, Togliatti, 2003.

[21] J.H. Friedman, On Bias, Variance, 0/1—Loss, and the Curse-of-Dimensionality, Data Mining and Knowledge Discovery 1, 55–77 (1997). https://doi.org/10.1023/A:1009778005914.

[22] F. Herrera, F. Charte, A.J. Rivera, M.J. del Jesus, Multilabel Classification Problem Analysis, Metrics and Techniques, Springer International Publishing, Switzerland, 2016. doi: 10.1007/978-3-319-41111-8.

[23] Kuznietsova N. V. Analytical Technologies for Clients' Preferences Analyzing with Incomplete Data Recovering, in: CEUR Workshop Proceeding, 2018, pp.118-128. URL: http://ceur-ws.org/Vol-2318/.

[24] D. Kwiatkowski, P. Phillips, P. Schmidt, Y. Shin, Testing the null hypothesis of stationarity against the alternative of a unit root, Journal of Economics, n. 54 (1992) 159–178.

[25] N. V. Kuznietsova, P. I. Bidyuk, Theory and practice of financial risk analysis: systemic approach, Lira-K, Kyiv, 2020.

[26] M. Hossin, M.N, Sulaiman, A Review on Evaluation Metrics for Data Classification Evaluations, International Journal of Data Mining & Knowledge Management Process 5 (2015) 1-11. doi: 10.5121/ijdkp.2015.5201.