sciendo

NS global economics ECONOMICS
Volume 10, No. 1, 2022

www.economicsrs.com

FORECASTING THE BEHAVIOR OF TARGET SEGMENTS TO ACTIVATE ADVERTISING TOOLS: CASE OF MOBILE OPERATOR VODAFONE UKRAINE

Tetiana Zatonatska[1], Oleksandr Dluhopolskyi[2,3], Tatiana Artyukh[1], Kateryna Tymchenko[1]

Received 15. 03. 2022. | Sent to review 18. 03. 2022. | Accepted 17. 05. 2022.

Original Article

[1] Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

[2] Higher School of Economics and Innovation (WSEI), Lublin, Poland

[3] West Ukrainian National University, Ternopil, Ukraine

Corresponding Author:
Oleksandr Dluhopolskyi
E-mail: dlugopolsky77@gmail.com

JEL Classification:
D22, D47, M21, M31, M37

Doi: 10.2478/eoik-2022-0005

UDK: 621.396.21;316.776.1(477)

ABSTRACT

Today, the use of machine learning technology in combination with the use of big data are topics that are actively discussed in business around the world. This topic has long gone beyond the information sphere, as it now applies to almost every sphere of life: economic, telecommunications, education, medicine, administration, and especially defense. Predicting customer behavior based on scoring models is in its infancy in Ukrainian companies, the main ones being the introduction of artificial intelligence technologies and machine learning, which will be the leading catalyst that will facilitate decision-making in business in the nearest future. The aim of the study is to develop a scoring model that predicts the behavior of target segments, namely, updating their activity to activate advertising tools. To achieve the goal of the work a set of research methods was used: dialectical – to reveal the theoretical foundations of models and types of forecasting models; analytical – in the study of the functioning of the environment SAS, Anaconda; optimization methods – to choose the best model and generate features. Scientific novelty and theoretical significance lie in the development of a scoring model for predicting the activity of subscribers of the telecommunications company "VF Ukraine", on the basis of which marketing campaigns are conducted. With the help of the built-in scoring model, the company "VF Ukraine" can target its campaigns to retain subscribers. The marketing directorate of the enterprise can choose the TOP-20 or TOP-30 of the most prone subscribers to non-resumption of activity, i.e., tend to switch to other mobile operators, and hold promotions for them – providing additional gifts and bonuses, money to mobile account.

Keywords: *market segment, telecommunication sector, target, advertising, scoring models, big data*

1. INTRODUCTION

Big data analysis methods have begun to find more and more practical application in the field of telecommunications to assess the level of customer satisfaction with their services, identify processes that caused dissatisfaction and predict which customers are going to change services and their providers. Although the industry sector is a leader in big data testing, especially due to the progress of Industry 4.0 (Dluhopolskyi et al., 2021), political actors are slower to engage in analytics (Zingale et al., 2018; Poel et al., 2018; Washington, 2014; Stavytskyy et al., 2019), despite the level opportunities and preferences that digitalization provides to all, including to improve the process of implementing public policy based on data.

The active use of big data has become a necessity in decision-making in the XXI century, including due to the ever-increasing ability to create and assemble them with extraordinary speed. Thus,

the categories "big data" and "data analytics" have become fashionable over the past few decades, which has led to their active use in various spheres of socio-economic life (Zhou et al., 2014; Zhou et al., 2020; Perkhofer et al., 2020; Polianovskyi et al., 2021). Billions of unstructured bytes of data in various locations need to be processed in order to obtain such information until relevant data is selected from them to build models. This type of analysis allows management to correct problematic processes or change people, as well as retain customers who are prone to changing suppliers (Machine Learning Mastery, 2015).

2. LITERATURE REVIEW

Many works of modern scientists are devoted to the relevance of the application of big data for the analysis of various processes. For example, G. Sekli and I. Vega, based on the evaluation of 256 respondents, investigate the factors that influence the adoption of big data analytics and evaluate the relationship it has with performance and knowledge management. This study provides practical guidance for decision-makers involved in or in charge of defining the implementation strategy of big data analytics in higher education institutions (Sekli, Vega, 2021).

Examples of the use of big data in telecommunications are described in detail in (Goworek, 2021), where the author substantiates the strategies of operators that improve competitive positions in the market, attract a wide customer base using big data. Authors (Radukić et al., 2019) consider the effects of digital transformation in the telecommunications markets which are characterized by network externalities. Also, in the article present a methodological framework for measuring the information society. In the study (Yusuf-Asaju et al., 2017) the authors propose to apply the process of estimating perceived quality of experience to assist mobile network operators to effectively manage network performance and aid satisfactory provision of mobile internet services. In the study (Simaković et al., 2021) the authors propose a big data solution that can collect and process huge amounts of data to extract valuable information and help mobile operators to bring their networks to enhanced quality level and offer full IoT solutions to their customers.

The study of X. Dai emphasizes the importance of user feedback, which is an excellent source of information for mobile operators. The author used several sets of data for the experiment – one is related to complaints, the other to survey data (Dai, 2017). The research of Ch. Dang focuses on two kinds of targeting in the mobile industry: to target churning customers and to target potential customers (Dang, 2017).

In the article (Varga, Gabor, 2021) the authors proved that with the help of marketing on social networks companies can easily advertise and promote products and services, targeting relevant segments. The main hypothesis of the author's research, which was confirmed, was that social networks influence people in their travel decisions.

In the study (Truong et al., 2017), the authors explore web analytics tools that explore ways to help entrepreneurs collect and process data from their websites to improve an organization's business strategy. Google and Woopra analytics tools are used to demonstrate how web analytics tools can help making proper adjustment to the business.

In the article (Rosario et al., 2021) it is proved that marketing approaches, such as IoT, big data, artificial intelligence, and machine learning, allow tracking in real time the relationship between business efficiency and advertising.

Many studies focus on the importance of big data in the e-commerce and e-business industry (Riddle, 2020; Ryfiak, 2020; Zatonatska et al., 2019; Zatonatska et al., 2021; Mykhalchuk et al., 2021) or cryptocurrency market development (Suslenko et al., 2022; Zatonatska et al., 2022). Such research is extremely relevant for forecasting and making optimal management decisions in times of uncertainty and risk.

3. RESEARCH OBJECTIVES

In our research, we will focus on using big data to forecast the demand for mobile operator services (Vodafone case) and prevent the outflow of customers to other mobile operators. The main research objectives include:

1. analyze the activities of the telecommunication company "VF Ukraine" and build a scor ing model that predicts subscriber behavior;
2. analyze the conditions under which there may be heterogeneity between target and non-target groups on the basis of recovery of activity;
3. predict the behavior of target segments to activate advertising tools.

4. METHODS

To predict the behavior of target segments, we conducted modeling of big data. To build the model data was processed to convert it from its "raw state" to the format used in machine learning algorithms. Initially, all column values were summarized by month (average, number of values, sum, maximum, minimum) for each numeric column for each client, and for category columns, different values were calculated. The type of functions was computed based on the social activity of subscribers via SMS and calls. The SAS software product is used in the research for both statistical and social characteristics.

Statistical functions are generated from all types of CDRs, such as the average number of calls made by a customer per month, the average number of Internet connections, the number of prepaid packages, the ratio of calls to SMS messages and many functions derived from CDR data aggregation.

Because we have data on all actions of network subscribers, they were grouped into groups related to calls, SMS, and Internet usage for each customer on a daily, weekly, and monthly basis for each action for nine months. Therefore, the number of generated functions is three times the number of columns. In addition, features related to complaints from customers from all systems have been introduced. Some features were related to the number of complaints, the percentage of complaints covered for all complaints, the average duration between each two complaints in a row, the duration in hours until the complaint was closed, the outcome of the closure and other features. IMEI-related features such as device type, brand, dual or single device, and number of devices changed by the client have been removed.

In addition, several functions have been created, such as the percentage of incoming / outgoing calls, SMS to competitors and landlines, binary functions that show whether customers subscribe to certain services or not, Internet speed between 2G, 3G and 4G, the number of devices used monthly, number of days out of coverage, percentage of friends related to a competitor, etc.

The main methods that have been used to analyze the behavior of target segments include (Hssina et al., 2014; Kuhn et al., 2013):

1. Venn diagram – detection of hidden relationship that allows you to combine multiple segments to identify connections, relationships, or differences. Its essence is to study customers who have purchased different categories of goods (for example, different pack ages of services), and on this basis to determine the possibility of cross-selling
2. Data profiling – defining client attributes when selecting records from a particular data tree, then based on such profiles generating subscriber profiles, which indicate their gen eral features and behavior. These profiles are then used to form an effective promotion and sales strategy;

3. Forecasting – analysis of time series, which allows you to adapt to changes, trends, and seasonal fluctuations. With forecasting, you can accurately predict monthly sales or the number of expected orders in any given month;
4. Mapping – defining geographical areas, using color coding to determine customer behavior when changing positions between geographic regions. The map, divided into polygons representing geographical regions, shows where subscribers are concentrated and where specific products are sold the most;
5. Rules of association – analysis of the consumer basket (cause / effect). This technique identifies relationships or affinity structures in the data and forms sets of rules that allow you to automatically select the rules that are most useful for key business perceptions (e.g., Which products would you buy at the same time and when? Which services customers you wouldn't buy and why? Which new cross-selling opportunities exist?);
6. Decision tree – classification and prediction of behavior, which is one of the most popular methods of classification in various data mining programs. Classification helps to perform operations such as choosing the right products that will be recommended to specific customers and predicting potential feedback. The most used decision tree algorithms are: ID3, C4.5 and CART.

The data operations mentioned above are performed using the following software products (Hastie et al., 2017; Hardikar et al., 2012; McLachlan et al., 2004; White, 2015):

1. Apache Hadoop – open-source software that consists of two main components: a distributed MapReduce processing system and a distributed Hadoop file system (HDFS). One of the most important reasons for using this program is the ability to process and analyze large amounts of data, which is impossible in many other systems. Storage is provided by HDFS, and analysis is performed by MapReduce. Although Hadoop is known for its MapReduce feature and its distributed file system, other subprojects provide additional services and build a foundation for providing high-level abstractions.
2. The Hadoop Distributed File System (HDFS) – memory component that offers a distributed architecture for extremely large-scale storage that can be easily expanded if needed. When a file is saved in HDFS format, it is divided into blocks of the same size. The block size can be adjusted or used predefined. The customer dataset is stored in HDFS format and contains many customer records of purchases made. The source file containing the decision rules is also written to HDFS.
3. MapReduce – programming model for processing and generating large data sets with a parallel distributed algorithm on a cluster. MapReduce works by dividing processing into two phases: the map phase and the reduction phase. In each phase there is a pair of keys and values of inputs and outputs, the types of which can be selected by the analyst. The analyst also defines two functions: the Map function and the Reduce function. Entry into the map phase is raw customer data. The result of the map function is processed by the MapReduce function before being sent to the reduce function. This processing sorts and groups key-value pairs by key.

The logic of model programming for the respective stages of the analysis is based on the C4.5 algorithm. It is an algorithm to generate a decision tree. Solution trees generated by C4.5 can be used for classification (which is why C4.5 is often called a statistical classifier). C4.5 uses information gain as a criterion for splitting. It can accept data with categorical or numeric values. To process continuous values, it generates a threshold and then divides the attributes with values above the threshold and values equal to or below the threshold. The C4.5 algorithm can easily handle missing

values because missing attribute values are not used in C4.5 gain calculations.
The flow of the system is shown at Fig. 1 (Deroos et al., 2014):

1. load the client data set with HDFS as input for the algorithm;
2. calling an instance of class C4.5;
3. using the MapReduce Hadoop structure, call the Map function, which checks whether this instance belongs to the current node or not. For all uncovered attributes, it displays the index and its value and the instance class label;
4. reduce function counts the number of matching cases (index and its value and label class);
5. entropy, information growth factor and attribute gain factor are calculated;
6. processing of the input data set with HDFS according to the defined algorithm for outputting C4.5 data from the decision tree within MapReduce;
7. create a decision rule and save it in HDFS format;
8. receiving new test data from the web interface;
9. gaining access to the rules and determining the categories of new data based on them;
10. providing visualization of the HDFS data set in the web interface in the form of histograms, pie charts, etc.

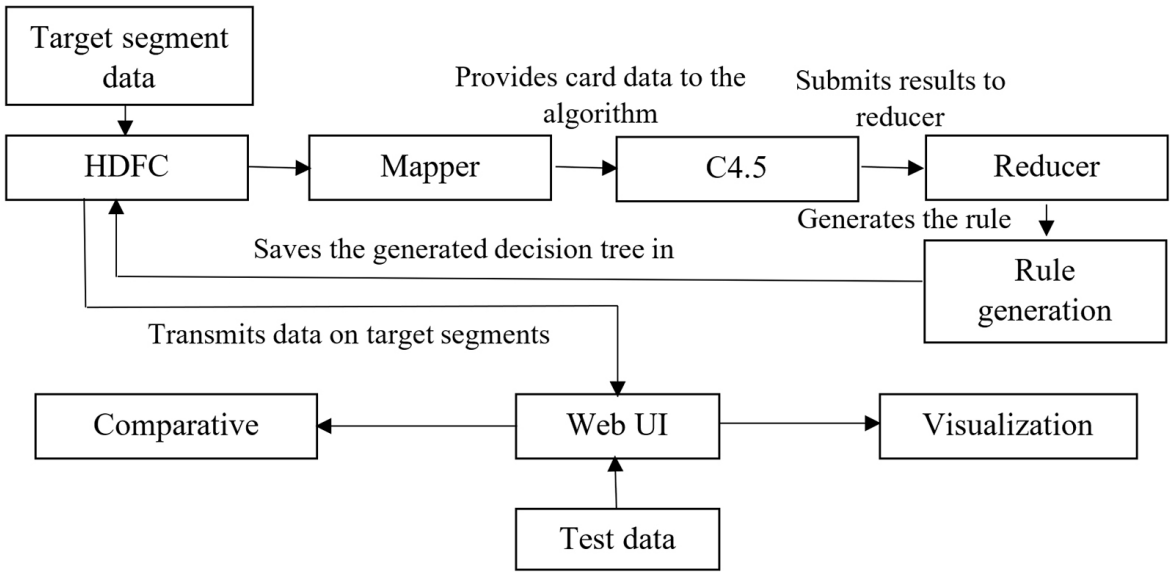
After data preparation and model training, the efficacy of the developed model is calculated using certain evaluation methods.
The first method of evaluation is the error matrix (confusion matrix) and other methods based on it. The error matrix is a tool for determining the performance of the classifier. It contains information on actual and estimated classifications. Fig. 2 shows the error matrix of the two-class, spam, and non-spam classifier.
There are four types of outputs in the error matrix:

1. True Positive (TP) – the number of correct predictions of an example being positive, which means that the positive class is correctly defined as positive;
2. False Negative (FN) – the number of incorrect predictions that the example is negative, which means that the positive class is incorrectly identified as negative;
3. False Positive (FP) – the number of incorrect predictions that are positive, which means that the negative class was incorrectly defined as positive;
4. True Negative (TN) – the number of correct predictions that the example is negative, which means that the negative class was correctly identified as negative.

The F1 score – is the weighted average of the response (sensitivity) and accuracy. F1 score allows you to assess the balance of the model, as well as its accuracy and response:

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \quad (1)$$

Figure 1. The flow of the system C4.5.



Source: (White, 2015; McLachlan et al., 2004).

Figure 2. Error matrix.

Actual class \ Predicted class	**Positive**	**Negative**	
Positive	*True Positive (TP)*	*False Negative (FN)* Type II Error	$Sensitivity \frac{TP}{(TP+FN)}$
Negative	*False Positive (FP)* Type I Error	*True Negative (TN)*	$Specificity \frac{TN}{(TN+FP)}$
	$Precision \frac{TP}{(TP+FP)}$	$Negative\ Predictive\ Value \frac{TN}{(TN+FN)}$	$Accuracy \frac{TP+TN}{(TP+TN+FN+FP)}$

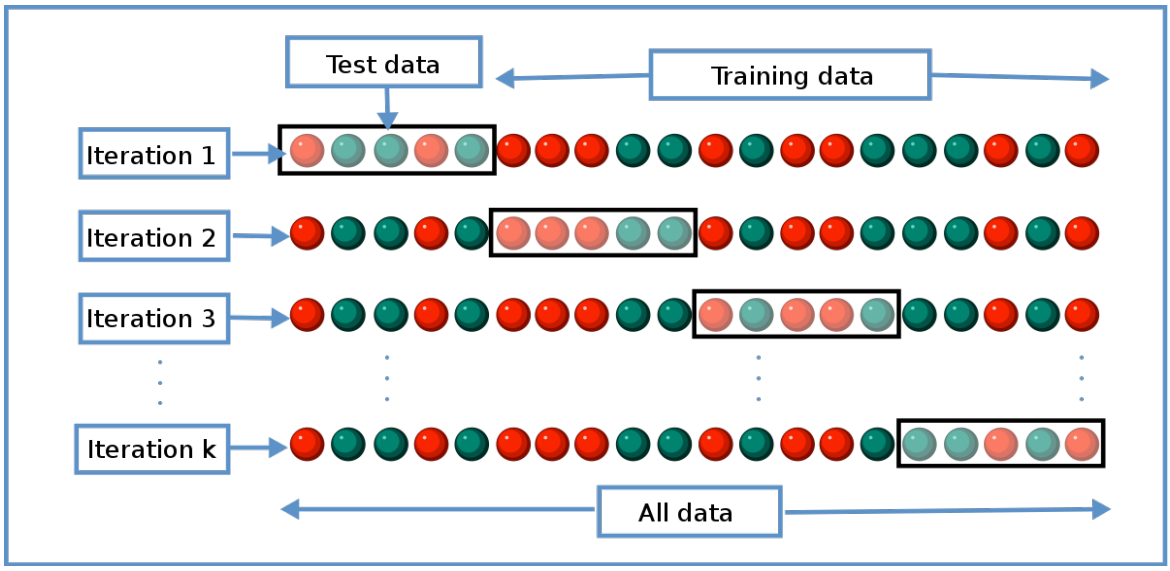
Source: (Lurie, 2014; Fawcett, 2006).

The above estimate evaluation helps to calculate the response and accuracy in one equation, and solves the problem of distinguishing models with low response and high accuracy or vice versa (Lurie, 2014).

The second method of evaluation is cross-validation. The main problem it allows to identify – the retraining of the model (Fig. 3). The definition of retraining is that the model is too specific for the datasets used and is likely to work very poorly for new observations (Celisse, 2014).

Another problem that can be identified through cross-validation is the lack of data. If the data is insufficiently processed, there is a risk of drawing general conclusions that basically under-use all the data. Because of these two problems – model retraining and data scarcity – the data set is usually divided into training and test sets. Then a learning algorithm (such as a decision tree) is applied to the training set, which is used to create a model. After that, the test set is submitted to the model that was taught using of the training set, and then performance is measured.

Figure 3. Cross-validation.



Source: (Celisse, 2014; Hastie et al., 2017; Fawcett, 2006).

Thus, the model is tested for processing the entire data set, which is the main idea of cross-checking. If incomplete data is processed, it is not used optimally. There are two main types of cross-validation – exhaustive and non-exhaustive (Airola et al., 2011; McLachlan et al., 2004; Molinaro et al., 2005; Hastie et al., 2017):

1. Exhaustive cross-validation methods are methods that study and test all possible ways of dividing the original sample into training and test kits. There are the following types of cross-validation:

1.1. The leave p-out cross-validation (LPO) involves the use of p-observations as a validation kit and the rest of the observations as a training kit. This is repeated in all ways of dividing the original set into a validation set and a training set. LPO cross-validation requires training and testing of the model times, where – the number of observations in the original sample, and – is the binary coefficient. For p > 1 and for moderately large n, cross-validation of LPO may be computationally impossible.

1.2. Leave-one-out cross-validation (LOOCV) is a special case of leave p-out cross-validation with p = 1. The process is like the previous one, but in cross-validation statistics are calculated on the remaining sample samples, while in LOOCV statistics are calculated only from saved samples. LOOCV requires less computational time than LPO cross-checking because there are only outputs, not . However, n outputs can still require a lot of time to calculate, and in this case, other approaches, such as cross-checking k-fold viewing, may be more convenient.

2. Non-exhaustive cross-validation methods do not calculate all the ways to divide the original sample that are possible. Therefore, they are only an approximation of exhaustive cross-validation and are divided into:

2.1. k-fold cross-validation, in which the original sample is randomly divided into k of the same size subsamples. Of the k subsamples, one is selected for model testing, and the remaining k-1 subsamples are used as training data. Then the process of cross-validation is repeated k times, and each of the k subsets is used exactly once as data for verification. Then k results can be averaged to obtain a single estimate. In stratified k-fold cross-validation, the number of subsamples is chosen so that the average response value is approximately equal in all subsamples. When cross-validation is repeated the data is randomly divided into k subsamples. Thus, the performance of the model can be averaged over several cycles.

2.2. Repeated random sub-sampling validation or Monte Carlo cross-validation method – creates several random splits of the data set into training and validation samples. For each such separation, the model is trained, and the accuracy of the prediction is assessed using validation data. Then the results are averaged over the number of divisions. The advantage of this method is that the share of training / validation splitting does not depend on the number of iterations (sections). The disadvantage of this method is that some observations can never be selected in a sub-sample for testing, while others can be selected more than once. In other words, test subsets can intersect. As the number of random splits approaches infinity, the result of re-checking the random subsample tends to the results of leave-p-out cross-validation (LPO). In the stratified version of this approach, random samples are generated so that the average response value (i.e., the dependent regression variable) is the same in the training and test sets.

5. RESULTS

Company "VF Ukraine", known to Ukrainians as the mobile operator Vodafone, is one of the largest players in the telecommunications market in Ukraine and ranks second in terms of market share among telecoms (market share 35%), after Kyivstar (market share around 48%) (Table 1, Fig. 4).
In 2019-2021, the company "VF Ukraine" actively developed new 4G databases in Ukraine, as well as conducted 5G testing.

Table 1. Main characteristics of key Ukrainian mobile operators, 2020.

№	Indicator	Kyivstar	Vodafone	Lifecell
1.	Market share 2021, %	48	35	17
2.	Number of subscribers, million	25,4	19,0	7,6
3.	Revenue, UAH billion	6	4,1	1,6
4.	Growth 2020/2019, %	6,8	9	7,7
5.	Earnings before interest, taxes, depreciation and amortization (EBITDA), UAH billion	4	…	0,9
6.	Operating income before depreciation and amortization (OIBDA), UAH billion	…	2,1	…
7.	Profitability, %	67,8	52,6	55,1
8.	Average revenue per unit (ARPU), UAH	72	68,4	70,3
9.	Volume of voice services, average number of minutes per subscriber	641	…	176
10.	Capital expenditures, UAH billion	1,7	1,9	0,4

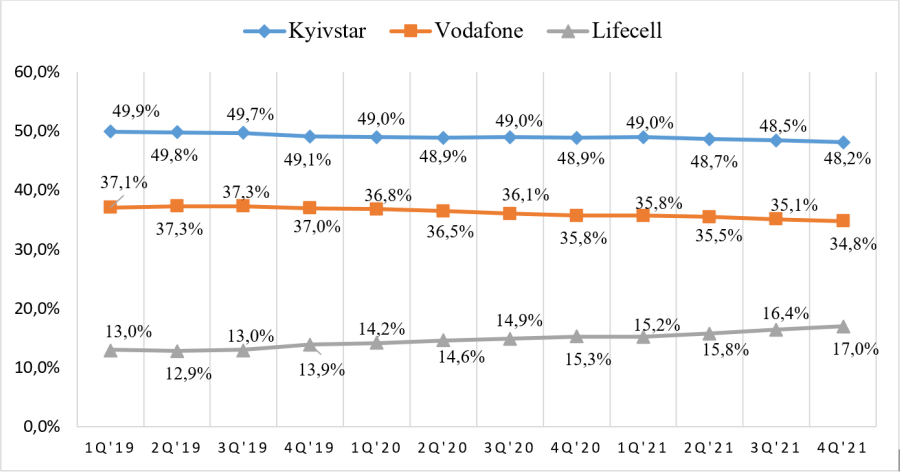
Source: developed by authors based on information from Kyivstar, Vodafone, Lifecell.

The key stage of building the model is the preparation of data based on the example of a telecommunications company ("VF Ukraine"). In order to build a scoring model that predicts subscriber behavior and answers the question "Will the subscriber resume activity within a week after four days of his last activity in the net?", it is necessary to collect historical data so that you can select homogeneously two groups of subscribers with the same behavior:

1. the first group, who did not display any activity for three days (weren't online, weren't making calls nor sending SMS messages) and did not replenish the account (these events were divided into two groups, because Vodafone recognizes them as different entities), and starting from the fourth day and during the week independently resumed their activity by performing any of the above events – the value of the target variable will be TARGET = 1;
2. the second group, who did not display any activity and did not replenish the account for three

days and starting from the fourth day and during the week did not resume their activity – the value of the target variable will be TARGET = 0.

Figure 4. Dynamic of the Kyivstar, Vodafone, and Lifecell market share, 2019-2021.



Source: developed by authors based on information from Kyivstar, Vodafone, Lifecell.

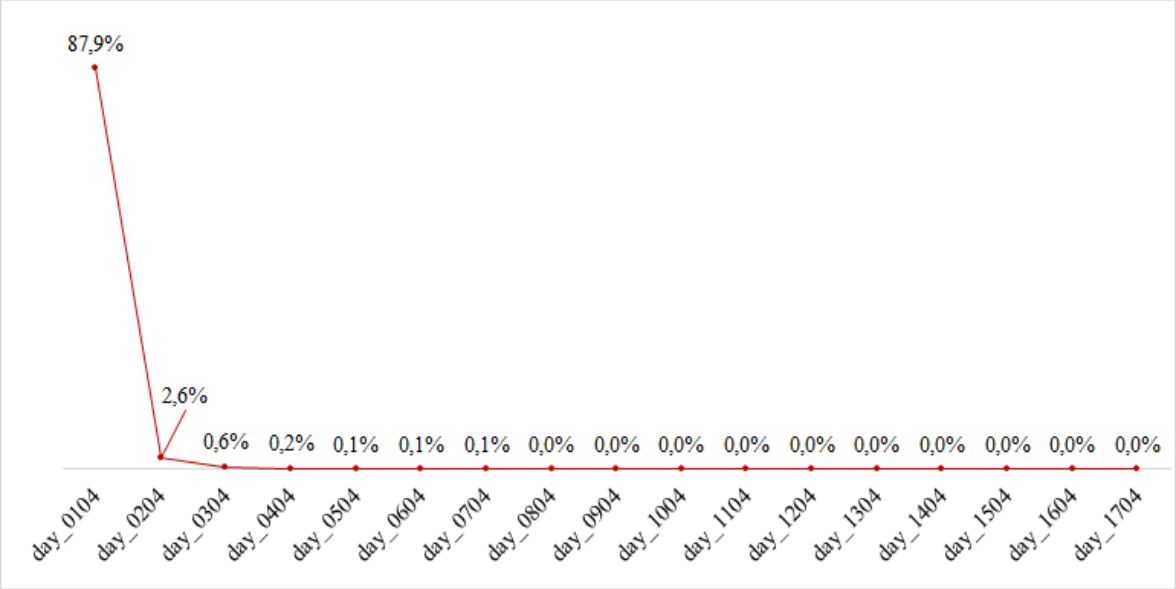
Therefore, the target of the model is to predict the resumption of subscribers' activity within a week after the third day when the subscriber did not perform any events.

The preconditions under which it was determined that it is necessary to make a forecast after the third day of inactivity was determined by a sharp decline in the percentage of activity recovery after the third day, as shown in Fig. 5. The graph shows how subscribers resume their activity. Active subscribers as of March 31, 2021, were taken as a base and was studied the dynamics of the next time they updated their activity.

In order to obtain homogeneity between groups, it was necessary to analyze the conditions under which there may be heterogeneity between target and non-target groups on the basis of recovery of activity. Thus, a period was chosen during which no promotions were held for subscribers – frozen zone (from 19.12.2020 to 13.01.2021), which allowed us to avoid the mistake of including users who didn't resume the activity on their own, but rather by motivation of the user to replenish the account and, as a result, the resumption of its activity through the accrual of bonuses to his mobile account because of bonuses in promotions in some subgroups.

Young users (subscribers who made their first event on Vodafone less than three months ago) were also removed from the prepared data. This is done as the company's policy in conducting marketing campaigns through CRM channels. This was done to obtain accurate data on subscriber activity within three months before the forecast date, as indicators such as the number of active days of new users may be underestimated because the subscriber simply has not yet appeared in the Vodafone network. The data was collected using the SQL programming language and the data step language built into the SAS software solution. All references to the tables refer to the internal libraries on the servers of "VF Ukraine".

Figure 5. Dynamics of subscribers out of inactivity.



Source: developed by authors based on information from Vodafone.

245 features were collected for 300,000 Vodafone subscribers, of which 50% of the target group: TARGET = 1 and the other 150,000 non-target subscribers (TARGET = 0). The features were selected in such a way that most of them are binary, and others – numerical. Time features have been converted to binary to increase the accuracy of the model. Missing values were also replaced with zeros.

Table 2 demonstrates the groups of features.

Table 2. Model's variables.

Group	Name	Description
Identifiers	app_n	unique subscriber number
	acc_n	personal account of the subscriber
Activity	act_w1_before	number of days of subscriber activity before the silence (from seven, per week of silence)
	act_w2_before	number of days of subscriber activity before the silence (from seven, a week before)
	act_w3_before	number of days of subscriber activity before the silence (from seven, the second week before)
	act_m_3	whether the subscriber had at least one day of activity per month (three months before)
	act_m_2	whether the subscriber had at least one day of activity per month (two months before)
	act_m_1	whether the subscriber had at least one day of activity per month (month before)
	act_w1_before_la13	number of days of subscriber activity before the silence, including incoming (from seven, a week before)
	act_w2_before_la13	number of days of subscriber activity before the silence, including incoming (from seven, the second week before)
	act_w3_before_la13	number of days of subscriber activity before the silence, including incoming (from seven, the third week before)
	Lifetime	number of days of the subscriber in the Vodafone network
	la_all_days_count	number of active days of the subscriber for the last month
	la_total_pauses	number of inactive days of the subscriber for the whole period
	la_total_active_days	number of active days of the subscriber for the whole period
	…	
Refill	mean_TU	average replenishment for the last 3 months
	last_TU_days	number of days from the last replenishment to silence
	last_TU_sum	the last amount of replenishment before the silence
	min_TU_dif	the minimum number of days between account replenishments
	…	
Use of the Internet	count_data_sessions_3d	number of date of sessions
	mean_data_dur_3d	average duration of the session date
	…	
SMS (number of SMS messages in groups)	sms_group_auto	number of incoming / outgoing SMS related to cars (refueling, repair, etc.)
	sms_group_shop	number of incoming / outgoing SMS related to stores
	sms_group_social_network	number of incoming / outgoing from social networks
	…	

Group	Name	Description
Device characteristics	device_days_usage	number of days of use of this phone
	device_brand_apple	whether there is an iPhone
	device_sim_count	number of SIM cards in the phone (whether this model supports two SIM cards)
	device_price	the price of the device
Calls	sum_call_cnt_out	number of outgoing calls for the last three months
	sum_call_cnt_in	number of incoming calls for the last three months
	sum_call_dur_out	the duration of all outgoing calls in the last three months
	sum_call_dur_in	the duration of all incoming calls in the last three months
Location	loc_cnt_events	number of events by location (for the last 3 months), where the subscriber made the largest number of events
	loc_is_obl_center	whether the location is a regional center
	loc_lat	latitude of the location
	loc_lon	longitude of the location
	loc_market_share	coverage of this region by Vodafone
	loc_is_nkt	whether the location is in an uncontrolled area
Participation in the real-time campaign	RTM_0015	was the subscriber a participant in the campaign promoting the "My Vodafone" application (for the last 3 months)
"Money to order" service	DNZ_COUNT_closed_loan_year1	number of closed debts for the service "Money to order"
	…	

Source: developed by authors based on (Vanwinckelen et al., 2012; Kuhn, Johnson, 2013; Vodafone).

Let's take a look at how subscribers were distributed in the target and non-target groups on the basis of activity in the "silence week". We can see a significant difference between these groups:

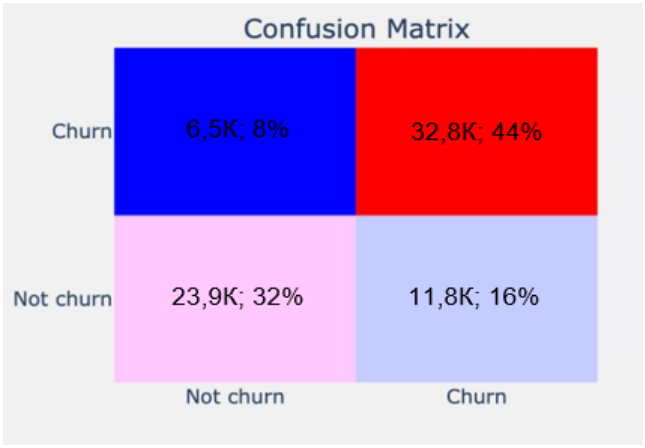
1. 21% of subscribers who resumed their activity in the next week were active only one day out of four in the week of silence (in the week when the subscriber did not resume activity for three days, we consider these days of silence, so the maximum possible active days in the last week – four), while not in the target group – as much as 38%;
2. 30% of subscribers from the target group were active all four days of the silence week, while in the group where subscribers did not resume activity the following week – 20%.

Let's also consider how subscribers were divided in the week to week of inactivity:

1. 30% of subscribers who did not resume their activity in the next week were not active on any day of the week before the week of silence, while only 11% of the target group of such subscribers;
2. from the target group 20% of subscribers were active for six days from seven a week before the week of silence, while in the non-target group – 13% of subscribers.

Let's build a logistic regression, which is a partial case of a linear classifier, in order to predict the probability of activity recovery of subscriber. First, we will develop a function that visualizes the results and will be universal to all algorithms for learning models. We will also set the division into training and test models.
As a result of learning the model, we obtain the results shown in Fig. 6.

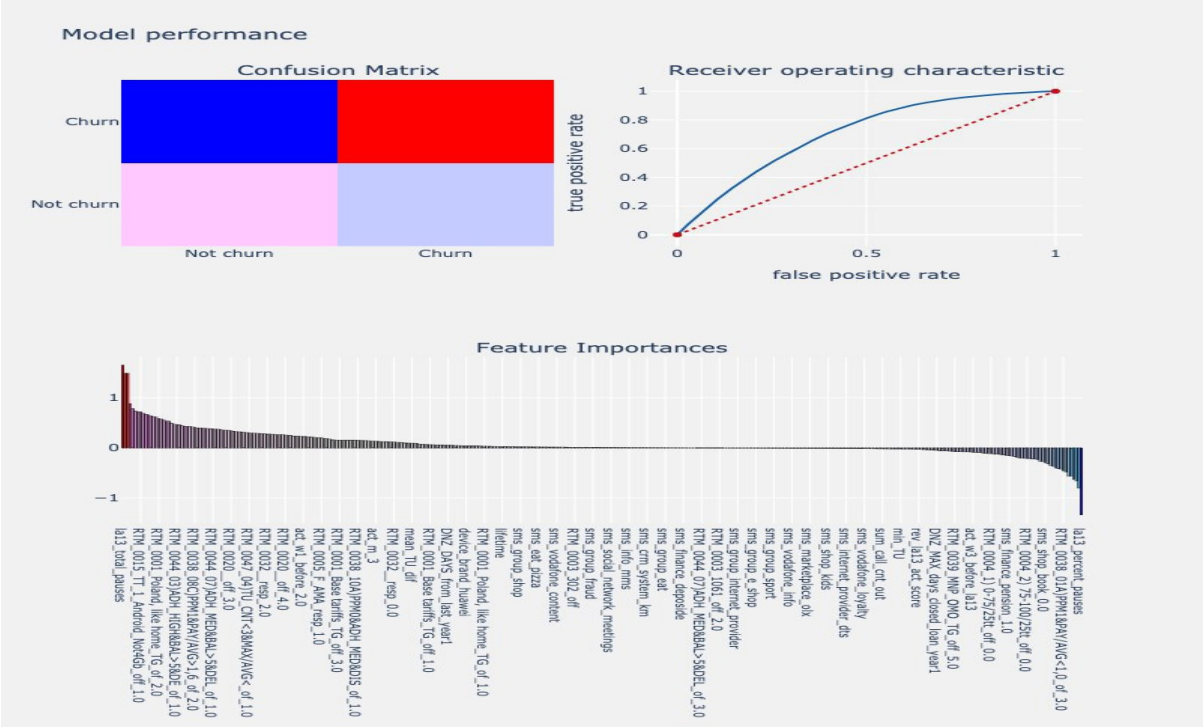
Figure 6. Logistic regression error matrix.



Source: developed by authors based on (Vanwinckelen et al., 2012; Kuhn, Johnson, 2013; Vodafone).

Next, we will show the most influential signs (Fig. 7).

Figure 7. Influence of logistic regression indicators.



Source: developed by authors based on (Vanwinckelen et al., 2012; Kuhn, Johnson, 2013; Vodafone).

The learning results of the model are shown in Fig. 8.

Figure 8. The results of learning the model by logistic regression.

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='ovr', n_jobs=1, penalty='l2', random_state=None,
                   solver='liblinear', tol=0.0001, verbose=0, warm_start=False)

 Classification report :
               precision    recall  f1-score   support

           0       0.79      0.66      0.72     34870
           1       0.73      0.86      0.79     35130

    accuracy                           0.76     75000
   macro avg       0.76      0.76      0.75     75000
weighted avg       0.76      0.76      0.75     75000

Accuracy   Score :  0.75654
Area under curve :  0.7563684653559225
```

Source: developed by authors based on (Vanwinckelen et al., 2012; Kuhn, Johnson, 2013; Vodafone).

6. DISCUSSION

The result of the study was the substantiation of theoretical and methodological principles and practical recommendations for predicting the behavior of target segments to activate advertising tools. The peculiarities of data preparation for forecasting in the telecommunications sector were studied. During the research, the main characteristics of subscribers for the period before forecasting the resumption of their activity were collected, segment analysis was performed, and a scoring model was built.

7. CONCLUSION

Based on statistical analysis and numerous experiments, the following conclusions were made:

1. collected and validated 245 of the most important characteristics for predicting the behavior of subscribers to restore their activity after three days of silence;
2. the main segments of subscribers, their distribution in target and non-target groups are anayzed;
3. few models of different machine learning algorithms are constructed (linear regression, linear regression with smoothing function, subtypes of decision tree algorithms: LightGBM and XGBoost);
4. based on the built models, the best in terms of accuracy, sensitivity, f1-score was selected, which was obtained using the LightGBM algorithm with an accuracy score of 0.78.

With the help of the built-in scoring model, the company "VF Ukraine" can target its campaigns to retain subscribers. The marketing directorate of the enterprise can choose the TOP-20 or TOP-30 of subscribers most prone to ceasing activity (tend to switch to other mobile operators), and conduct maintenance campaigns for them – providing additional gifts and bonuses, money to mobile account, etc. In the future, this model can be improved with the technical capabilities of the company to obtain more data about subscribers, and therefore, they can be included in the analysis to improve the accuracy of forecasts.

REFERENCES

Airola, A., Pahikkala, T., Waegeman, W., De Baets, B., & Salakoski, T. (2011). An experimental comparison of cross-validation techniques for estimating the area under the ROC-curve. *Computational Statistics & Data Analysis*, 55(4), 1828-1844. https://doi.org/10.1016/j.csda.2010.11.018

Celisse, A. (2014). Optimal cross-validation in density estimation with the L^2-loss. *The Annals of Statistics*, 42(5), 1879-1910. https://doi.org/10.1214/14-AOS1240

Dai, X. (2017). Identifying dissatisfied 4G customers from network indicators: a comparison between complaint and survey data. *Big Data Applications in the Telecommunications Industry*, 41-53. https://doi.org/10.4018/978-1-5225-1750-4.ch004

Dang, Ch. (2017). Network-based targeting: Big Data application in mobile industry. *Big Data Applications in the Telecommunications Industry*, 78-107. https://doi.org/10.4018/978-1-5225-1750-4.ch007

Deroos, D., Zikopoulos, P.C., Melnyk, R.B., Brown, B., & Coss, R. (2014). Hadoop for dummies. John Wiley & Sons, Inc., Hoboken, New Jersey. ISBN-13:978-1118607558

Dluhopolskyi, O., Simakhova, A., Zatonatska, T., Oleksiv, I., & Kozlovskyi, S. (2021). Potential of virtual reality in the current digital society: economic perspectives. 11th International Conference on Advanced Computer Information Technologies (September 15-17, 2021). Deggendorf, Germany, 360-363. https://doi.org/10.1109/ACIT52158.2021.9548495

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Goworek, K. (2021). The big impact of Big Data on the telecom industry. https://tasil.com/insights/big-data-in-telecoms.

Hastie, T., Tibshirani, R., & Friedman, J. (2017). Elements of Statistical Learning: data mining, inference, and prediction. 2nd Edition. https://hastie.su.domains/ElemStatLearn.

Hssina, B., Merbouha, A., Ezzikouri, H., Erritali, M. (2014). A comparative study of decision tree ID3 and C4.5. International Journal of Advanced Computer Science and Applications (IJACSA), Special Issue on Advances in Vehicular Ad Hoc Networking and Applications, 13-19. https://doi.org/10.14569/SPECIALISSUE.2014.040203

Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. New York: Springer. https://link.springer.com/book/10.1007/978-1-4614-6849-3.

Lurie, A. (2014). 39 Data Visualization Tools for Big Data. Profit Bricks, The Laas Company. https://cloud.ionos.com/compute.

Machine Learning Mastery (2015). Discover Feature Engineering, How to Engineer Features and How to Get Good at It. https://machinelearningmastery.com.

McLachlan, G.J., Do, K.-A., & Ambroise, C. (2004). Analyzing microarray gene expression data. Wiley. ISBN: 978-0-471-72842-9

Molinaro, A.M., Simon, R., & Pfeiffer, R.M. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21(15), 3301-3307. https://doi.org/10.1093/bioinformatics/bti499

Mykhalchuk, T., Zatonatska, T., Dluhopolskyi, O., Zhukovska, A., Dluhopolska, T., Liakhovych, L. (2021). Development of recommendation system in e-commerce using emotional analysis and machine learning methods. The 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS). Vol.1 (September 22-25, 2021). Cracow, Poland, 527-535. https://ieeexplore.ieee.org/document/9660854

Perkhofer, L., Walchshofer, C., & Hofer, P. (2020). Does design matter when visualizing Big Data? An empirical study to investigate the effect of visualization type and interaction use. *Journal of Management Control*, 31, 55-95. https://doi.org/10.1007/s00187-020-00294-0

Poel, M., Meyer, E.T., & Schroeder, R. (2018). Big data for policymaking: great expectations, but with limited progress? *Policy & Internet*, 10(3), 347-367. https://doi.org/10.1002/poi3.176

Polianovskyi, H., Zatonatska, T., Dluhopolskyi, O., & Liutyi, I. (2021). Digital and technological support of distance learning at universities under COVID-19 (case of Ukraine). *Revista Romaneasca pentru Educatie Multidimensionala*, 13(4), 595-613. https://doi.org/10.18662/rrem/13.4/500

Radukić, S., Mastilo, Z., & Kostić, Z. (2019). Effects of digital transformation and network externalities in the telecommunication markets. *ECONOMICS*, 7(2), 31-42. https://doi.org/10.2478/eoik-2019-0019

Riddle, J. (2020). How Will Big Data Transform E-Commerce Marketplaces? https://learn.g2.com/big-data-ecommerce.

Rosario, A., Moniz, L.B., & Cruz, R. (2021). Data science applied to marketing: a literature review. *Journal of Information Science and Engineering*, 37(5), 1067-1081. https://doi.org/10.6688/JISE.202109_37(5).0006

Ryfiak, S. (2020). Big Data is taking eCommerce by storm. Here's why you can't wait it out. https://www.business2community.com/ecommerce.

Sekli, G.F., & Vega, I. (2021). Adoption of Big Data analytics and its impact on organizational performance in higher education mediated by knowledge management. *Journal of Open Innovation: Technology, Market, and Complexity*, 7(4), 221. https://doi.org/10.3390/joitmc7040221

Simaković, M.N., Cica, Z.G., & Masnikosa, I.B. (2021). Big Data architecture for mobile network operators. 15th International Conference on Advanced Technologies, Systems and Services in Telecommunications (TELSIKS), 283-286. https://doi.org/10.1109/TELSIKS52058.2021.9606290

Stavytskyy, A., Dluhopolskyi, O., Kharlamova, G., Karpuk, A., Osetskyi, V. (2019). Testing the fruitfulness of the institutional environment for the development of innovative-entrepreneurial universities in Ukraine. *Problems and Perspectives in Management,* 17(4), 274-288. http://dx.doi.org/10.21511/ppm.17(4).2019.23

Suominen, A., & Hajikhani, A. (2021). Research themes in big data analytics for policymaking: Insights from a mixed-methods systematic literature review. *Policy & Internet,* 13(4), 464-484. https://doi.org/10.1002/poi3.258

Suslenko, V., Zatonatska, T., Dluhopolskyi, O., Kuznyetsova, A. (2022). Use of crypto-currencies Bitcoin and Ethereum in the field of e-commerce: case study of Ukraine. *Financial and credit activity: problems of theory and practice,* 1(42), 62-72. https://fkd.net.ua/index.php/fkd/article/view/3603/3461

Truong, C., Phuong, H., Thi, N., & Trang, H. (2017). Web analytics tools and benefits for entrepreneurs. Bachelor's Thesis in Business Information Technology, 79 p. https://www.theseus.fi/bitstream/handle/10024/143135/Nguyen_Trang.pdf?sequence=2&isAllowed=y

Vanwinckelen, G., Blockeel, H., De Baets, B., Manderick, B., Rademaker, M., & Waegeman, W. (2012). On estimating model accuracy with repeated cross-validation. Proceedings of the 21st Belgian-Dutch Conference on Machine Learning, 39-44. ISBN: 978-94-6197-044-2

Varga, I.E., & Gabor, M.R. (2021). The influence of social networks in travel decisions. *ECONOMICS*, 9(2), 35-48. https://doi.org/10.2478/eoik-2021-0015

Washington, A.L. (2014). Government information policy in the era of big data. *Review of Policy Research*, 31(4), 319-325. https://doi.org/10.1111/ropr.12081

White, T. (2015). Hadoop: The Definitive Guide. O'Reilly Media, Inc. 4th Edition. ISBN: 9780596521974

Yusuf-Asaju, A.W., Dahalin, Z.B., & Ta'a, A. (2017). Mobile network quality of experience using big data analytics approach. 8th International Conference on Information Technology (ICIT) (May 17-18, 2017), 658-664. https://doi.org/10.1109/ICITECH.2017.8079923

Zatonatska, T., Dluhopolskyi, O., Chyrak, I., & Kotys, N. (2019). The internet and e-commerce diffusion in European countries (modeling at the example of Austria, Poland, and Ukraine). *Innovative Marketing*, 15(1), 66-75. http://dx.doi.org/10.21511/im.15(1).2019.06

Zatonatska, T., Fedirko, O., Dluhopolskyi, O., & Londar, S. (2021). The impact of e-commerce on he sustainable development: case of Ukraine, Poland, and Austria. IOP Conference Series: Earth and Environmental Science, Volume 915, (ISCES) "International Conference on Environmental Sustainability in Natural Resources Management" (October 15-16, 2021). Odesa, Ukraine. https://doi.org/10.1088/1755-1315/915/1/012023

Zatonatska, T., Suslenko, V., Dluhopolskyi, O., Brych, V., Dluhopolska, T. (2022). Investment models on centralized and decentralized cryptocurrency markets. *Naukovyi Visnyk Natsionalnoho Hirnychoho Universytetu*, 1, 177-182. https://doi.org/10.33271/nvngu/2022-1/177

Zhou, S., Zhang, X., Liu, J., Zhang, K., Zhao, Y. (2020). Exploring development of smart city research through perspectives of governance and information systems: A scientometric analysis using cite space. *Journal of Science and Technology Policy Management,* 11(4), 431-454. https://doi.org/10.1108/JSTPM-05-2019-0051

Zhou, Z.-H., Chawla, N.V., Jin, Y., & Williams, G.J. (2014). Big data opportunities and challenges: Discussions from data analytics perspectives [discussion forum]. *IEEE Computational Intelligence Magazine*, 9(4), 62-74. https://doi.org/10.1109/MCI.2014.2350953

Zingale, N.C., Cook, D., & Mazanec, M. (2018). Change calls upon public administrators to act, but in what way? Exploring administration as a platform for governance. *Administrative Theory & Praxis,* 40(3), 180-199. https://doi.org/10.1080/10841806.2018.1485447