Postmodern Openings
ISSN: 2068-0236 | e-ISSN: 2069-9387
Covered in: Web of Science (WOS); EBSCO; ERIH+; Google Scholar; Index Copernicus; Ideas RePeC; Econpapers; Socionet; CEEOL; Ulrich ProQuest; Cabell, Journalseek; Scipio; Philpapers; SHERPA/RoMEO repositories; KVK; WorldCat; CrossRef; CrossCheck

2022, Volume 13, Issue 1Sup1, pages: 78-104 | https://doi.org/10.18662/po/13.1Sup1/415
Submitted: January 9th, 2022 | *Accepted for publication:* February 5th, 2022

Customer Churn Prediction in Telecommunication Industry. A Data Analysis Techniques Approach

Denisa MELIAN[1],
Andreea DUMITRACHE[2],
Stelian STANCU[3],
Alexandra NASTU[4]

[1] Bucharest University of Economic Studies, Bucharest, Romania, denisa.melian@yahoo.com

[2] Bucharest University of Economic Studies, Bucharest, Romania, dumitrache.andreea03@gmail.com

[3] Bucharest University of Economic Studies, Bucharest, Romania, stelian_stancu@yahoo.com

[4] Bucharest University of Economic Studies, Bucharest, Romania, alenastu@yahoo.com

Abstract: Telecommunications is one of the most dynamic sectors in the market, where the customer base is an important pawn in receive safe revenues, so is important to focus attention is paid to maintaining them with an active status. Migrating customers from one network to another varies among telecommunication companies depending on different factors such as call quality, pricing plan, minute consumption, data, sms facilities, customer billing issues, etc. Determining an effective predictive model helps detect early warning signals when churn occurs and assigns to each customer a score called "churn score" that indicates the likelihood that the individual might migrate to another network over a predefined time period. To this extent, the present paper uses more than 10k customers sample of a telecommunication company and tries to analyse the churn behavior. The aim of the paper is both to test the efficiency and performance of the most commonly used data mining techniques to predict the churn behavior and to underline the main indicators that can be used when conducting such analyses. Knowing the magnitude of the churn phenomenon, the company can prevent the instability that is going to occur by applying a series of measure in order to increase the retention of the current customers.

Keywords: *data mining techniques; churn; customer's behavior; telecommunication.*

How to cite: Melian, D., Dumitrache, A., Stancu, S., & Nastu, A. (2022). Customer Churn Prediction in Telecommunication Industry. A Data Analysis Techniques Approach. *Postmodern Openings, 13*(1Sup1), 78-104. https://doi.org/10.18662/po/13.1Sup1/415

1. Introduction

The telecommunication industry is expanding its dynamism and competitiveness more and more every day. It continues to be the epicenter of growth and innovation for almost every industry on the market. Mobile devices and dependence on continuous connectivity in both the business and the employees are increasingly embedded in the structure of society, these being absolutely essential in stimulating the economic impetus around key trends.

Over time, this type of businesses in the telecommunications industry have multiple threats when we are talking about financial loss from customers who want to leave their actual telecom service provider and exchange it for other offers from other companies. A conclusive identification of the customers who might "leave" the network is beneficial for the telecommunications companies as they may have the change to propose a series of personalized offers, based on the customer's profile, preventing in this way their loss.

Due to the intense competition and the saturated market on which they operate, a lot of companies are aware that the database with existing customers is their best asset. This type trend is mostly important in subscription for postpaid segment. In general, this type of information are used to can build some type of models for churn prediction in the mobile industry that is based on customers demographics such as gender ,age or network age (the period of time in which a customer has been in a particular telecommunication network), contract dates, call details, complaint data, billing information, as Kotler and Keller (2016) and Hung et al. (2006) state.

Data science methods has proven applicability to predict churn behavior in many countries. According to Huang et al. (2012) neural networks and decision trees showed to be useful in managing telecommunications in Taiwan. The decision trees and logistic regression models has used for analyzes churn processes for a specific data set received from a UK mobile operator. In this type of model's evaluation section, it has been shown that this type of classification (trees) exceeded the classic logistic regression. The customer response prediction for US telecoms companies was performed with this type of techniques (artificial neural networks by Tsai & Lu, 2009).

Research results have shown that support vector machines (SVMs) are also very useful for churn prediction, showing best performance for Naive Bayes classifier, artificial networks and decision trees, according to Zhao et al. (2005). Even more, Kisioglu and Topcu (2011) underlined that Bayesian-based classification has identified factors that affect churn in the telecommunications industry in Turkey. Nevertheless, a special interest is related for improving the predictive performance of these type of classifiers. In this context, De Caigny et al. (2018) believe that one of the best solutions in this case might be the creation of a hybrid algorithm that is based on a logistic regression and some decision trees.

In this context, the present paper addresses this widely discussed topic, the prediction of churn on a dataset of a major telecommunication company in Romania. The paper analyzes the behavior of the customers in a telecommunication company through five statistical methods of modeling and prediction, having as main objective the selection of the variables that might give a "hint" on the individuals who might give up the services offered by the company analyzed for other telecommunication services from a competing company. The analysis is developed for postpaid customers as, according to a report provided by the National Authority for Administration and Regulation in Communications (2016), in the end of 2016 year in Romania existed 22.9 million active clients (telecom industry), but 11.5 million were postpaid customers. A sample of 10715 customers provided by a telecommunication company is used in analysis.

2. Literature review

Companies are beginning to change their classic marketing strategies for "targeted" marketing as they have seen over time that it is profitable to keep satisfied existing customers than constantly reach new customers, described by a high rate of lack of knowledge regarding their profiling and history. The highest priority in this process is the identification of those customers prone to change.

As for the mobile telecommunication industry, the data used to buil churn prediction models includes the following aspects: contact data, customer history and characteristics (such as: age, gender, the time spent in network), demographic information, call quality, complains, bille and payments (Wei & Chiu, 2002). Recently, for the prediction of churn in the field of telecommunication services, the specialists have proposed a set of variable features such as: service usage duration, billed amount, structure of monthly service charges and type of payment. There are a series of papers in the literature (Hung et al., 2006; Mozer et al., 2000; Wei & Chiu, 2002) dealing with how one should select aggregate call detail for predictors. As for the applications made within the literature, In the literature, churn prediction has been conducted based on the following combination of information (Euler, 2006; Hung et al., 2006; Mozer et al., 2000; Wei & Chiu, 2002) aggregate details about previous and current month calls and customer-specific details, such as customer address and payment method.

In order to achieve a predictive model, various techniques and methods analyzed in the literature have been proposed (Mozer et al., 2000; Wei & Chiu, 2002) as it follows: The Bayesian Classifier; The closest K neighbors; Logistic regression; Decision trees (DT); Artificial Neural Networks (ANN).

Acording to Mozer et al. (2000), of all the above-mentioned techniques, Decision Tree (DT) and Artificial Neural Networks (ANN) are the techniques that provide the best results for churn prediction. The autors also state that the application of the ANN technique to a database can obtain a prediction model with a higher accuracy than with the implementation of decision trees.

On the other hand, there are studies that state that the correct classification rate increases when decision trees are used (Qi et al., 2006). Even more, De Caigny et al. (2018) argue that this type of Support Vector Machines (SVM) would be more encouraging approach for loyalty customers (churn prediction), while Faris (2018) proposes a hybrid swarm intelligent Neural Network model.

2.1. Few words about the churn phenomenon

Churn in Telecommunications is the term used to describe collectively the cessation of customer service activity. Telecom operators make a hard competition to can keep their actual customers and require an efficient model for churn prediction to can monetarize the actual customer situation. Also, Faris (2018) states that in the telecommunication market, it is easy for the customers to end their subscription and switch to other companies. The author underlines even the fact that the cost of acquiring new customers is higher than retaining the existing ones (Faris, 2018). In this context, the prediction of this action has become an important part of the strategic making decisions and business plan process in telecommunications, so it is very important to anticipate customer behavior.

According to Huang et al. (2012) , churn customers can be classified in the following categories:

- Active churn (volunteer): those type of customers who want to cancel their actual contract to can move to another provider;
- Passive churn (non-volunteer): when your actual company whants to interrupts a customer service;

• Rotative churn (silent): this type of customers who have ended their actual contract without discussion between the two involved parties (client and company).

Active churners and passive churners can be easily predicted using traditional approaches. The problem arrives when one wants to evaluate the behavior of the third group of churners as this is difficult to predict as such types of customers can appear at any time, their behavior being unpredictable. New technologies that can be applied to the database not provide important information to the organization's decision department about past and current customers' behavior but can also provide future predictions using predictive modeling methods.

Also, according to Mattison (2005), from the point of view of the reasons behind this action, churn can be framed in the following typologies:

• Tariff Churn - those customers who migrate between packages offered by companies (for example, customers who move according to the tariff plans offered by different telecom companies);

• Qualitative Churn - customers migrating for better service quality (such as network quality and coverage at country level, stable service, a 4G coverage within a certain range);

• Family Churn - customers who have multiple connections on a network other than the current network.

To efficiently manage customers with a high churn level into a company, it is very important to can build an efficient and clean prediction model that acts as customer decision-maker support. To accomplish this, there are many modeling and prediction techniques that can best help in selecting customers who are most likely to quit current services. The implemented predictive model is influenced by the data mining techniques used, the selection method applied, the variables that are included in the model, and the time spent to can finish the final churn prediction model.

The churn forecast or the early identification of customers who can interrupt the use of a service is an important and increasingly important concern of many industries. Given that these businesses collect a growing amount of heterogeneous data on a large scale on customer characteristics and behaviors, it is possible to implement new churn prediction methods.

Customer churn rate is an indicator that measures the number of customers and subscribers who have stopped consuming a service provided by a company for a long time in return for a service provided by a competitor. The resulting value is translated as the percentage of customer abandonment or the loss of service abandonment. The simplest formula for measuring churn rate divides the number of subscribers lost in a month to the total percentage obtained earlier in the month. To reduce customers' churn rates, it is necessary to take into account several aspects of the business such as: Company image on the market; Timely identification of the reasons behind the churn process; Loyalty to customers; Fulfilling or adapting to customer expectations; Quality services and products; Segmentation of the database correctly.

The prediction of churn clients has a special interests for academic and industry, where numerous studies have been proposed for this process in various areas such as: financial service, banking, credit card accounts, airline, social, etc. Therefore, various churn prediction studies have been published in service-based industries. However, looking in the literature for the churn prediction modeling techniques, one can state that there is no standard model.

2.2. Data mining techniques used for anticipating the churn behaviour

The data mining process is an interdisciplinary computer science subdomain that helps us apply intelligent methods to extract data models. The general purpose of mining data is to extract information from a data set and transform it into an easy-to-understand structure for later use.

Data mining techniques are used in many types of areas, for example: retail, financial-banking, telecom, social networks, etc. The most important data mining techniques include classificators and prediction, grouping, association rules, sequence analysis, time series models as well as some new techniques such . In applicable applications used for real life, a data extraction process can be divided into six major steps: Understanding the necessity Knowing the variables; Preparing the data; Modeling; Evaluation; Model implementation.

As for the telecommunications industry, the most challenging problem faced is represented by the customer. A customer database flagged as predisposed to churn allows the company to target those customers and start applying different retention strategies to reduce the customer migration percentage.

3. Customer churn prediction in telecommunications

In this study, we aim to analyze and predict the churn phenomenon by considering one of the largest mobile network operators in Romania. This operator has approximately five million active subscribers and has agreed to provide access to an anonymized database, which contains historical data related to its customers.

3.1. Data gathering and data analysis

A sample of 10715 subscribers have been randomly selected from the provided database. The churn clients are labeled with 1 in the churn variable, indicating whether the client in the analyzed data set migrated to other mobile carrier companies 3 months after the baseline query or he/she is still an active client. The number of churn customers is 1468 individuals, representing 13.70% among all the customers included in the sample.

The following data have been extracted for each customer:

- Demographic data: area , age and gender ;

- Information about the clients lifecycle: tenure (Tenure), company in age (period in months since he/she is in the network), date of network telephone activation (Data), number of years/months since the client change of the last offer on account (MonthsO), number of months from his request through telesales to modify the actual subscription offer for another service - change of the current contract (MonthsC), the type of the last contract between the company and the client - the type of offer he or she has installed (Contract);

- Data on the financial strength - this category contains details about the average for a minimum three months electronic invoice that each customer is necessary to pay, the value for the extra cost for services (ExtraCosts);

- Subscriber actions with competing telecom clients: the sum for national minutes and assumed synthesized in two type of variables (MinC and MinR).

Considering the demographic data, it can be observed that 12.67% are located in the rural area (representing 1358 persons), while 87.33% are from the urban area (9357 persons). Regarding the gender, 4374 persons are female (40.82%), while 6341 are male (59.18%). Also, 4790 persons had no extra costs paid for off-site services (44.70%), 5610 persons had a cost smaller than 10 euros (52.36%), while 315 persons has a cost larger than 10 euros (2.94%) – Figure 1.

Based on the invoice’s value it has been determined that for 7269 persons, it has been smaller than 10 euros, for 3293 persons the invoice has been between 10 and 20 euros, while 153 persons has been more than 20 euros.

As for the subscriber interaction with competing network clients, it has been observed that for the MinR variable (received minutes), there are 360 persons who have no minute received from another person in a competing network. Looking further on the 360 persons, it has also been observed that none of these persons have initiated a call to a person having a phone number in a competing network. As for the rest of 10355 persons, 4225 persons have received less than 30 minutes (39.43%), 4073 persons have received calls totaling between 30 and 60 minutes (38.01%), 1737 persons have had total calls between 60 and 90 minutes (16.21%), while 320 persons have received more than 90 minutes (2.99%). On the other hand, considering the number of minutes made to persons having the number in competing networks (MinC variable), 466 persons have never initiated a call in another network (4.35%), 2774 persons have used less than 30 minutes in other networks (25.89%), 1843 persons have used between 30 minutes and 60 minutes (11.37%), while the great majority. 4414 persons have used more than 90 minutes in a foreign network (41.19%).

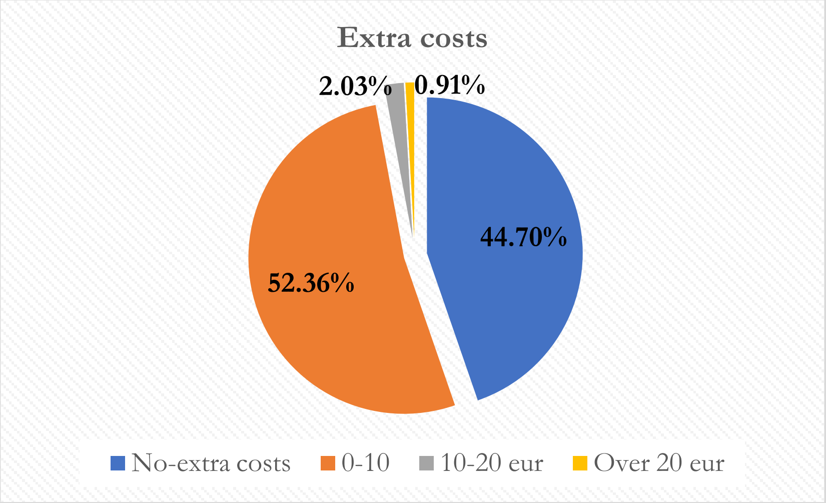


Figure 1. Percentage of persons with and without extra costs

For the tenure, for 79.92% of the persons a tenure between 145 and 199 has been recorded, while for the rest of 20.08%, the tenure has been between 200 and 241. When we are talking about the number of months since the last change of the offer on the actual account, the data in Figure 2 has been encountered. It can be observed that 44.28% of the persons have a contract with an age between 1 and 2 years, 32.05% have a contract with an age less than 1 year, 15.60% between 2 and 3 years and 8.06% a contract with a data greater than 3 years.

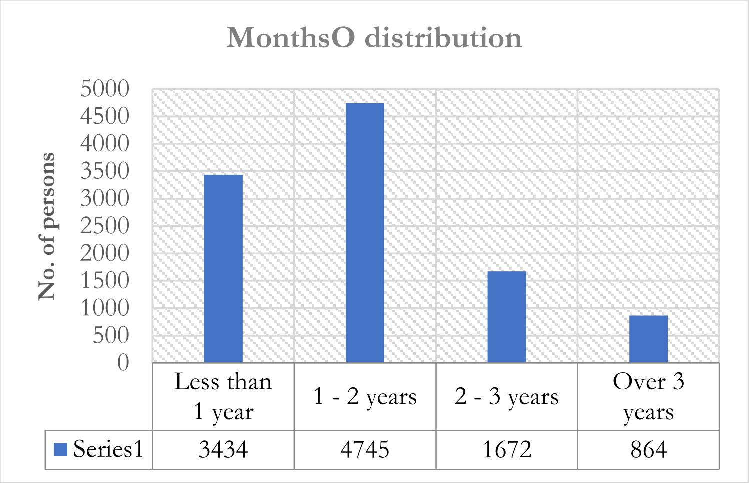


Figure 2. Person's distribution over the considered categories for MonthsO variable

As for the MonthsC variable, 8814 persons have a renewed contract in the last 12 months (82.26%), while the rest, have a renewed contract older than 12 months (17.74%).

3.2. Analysis steps

In order to determine the indicators which can best underline the churn behavior of the considered sample, the steps presented in Figure 3 have been taken. As it can be observed the analysis is divided into two main directions. The first one is represented by applying the k-means clustering and dividing the set into several clusters on which we have further applied logistic regression and decision tree. The second one is represented by the use of Random Forest and Balanced Random Forest on the whole data sample. Knowing the best descriptive variables for the churn behavior, then, we have applied PSM in order to explaining the net effect produced by these over the churn behavior.

In addition to analyzing the churn behavior, we also aim to test the efficiency and performance of the most commonly used data mining techniques to predict the churn behaviour. The purpose of conducting this analysis is to detect in a timely manner the best method that can be used by the companies in order to detect the churn behavior of their customers. Knowing the magnitude of the churn phenomenon, the company can prevent the instability that is going to occur by applying a series of measure in order to increase the retention of the current customers. Therefore, the data mining techniques presented in the previous chapter will be applied.

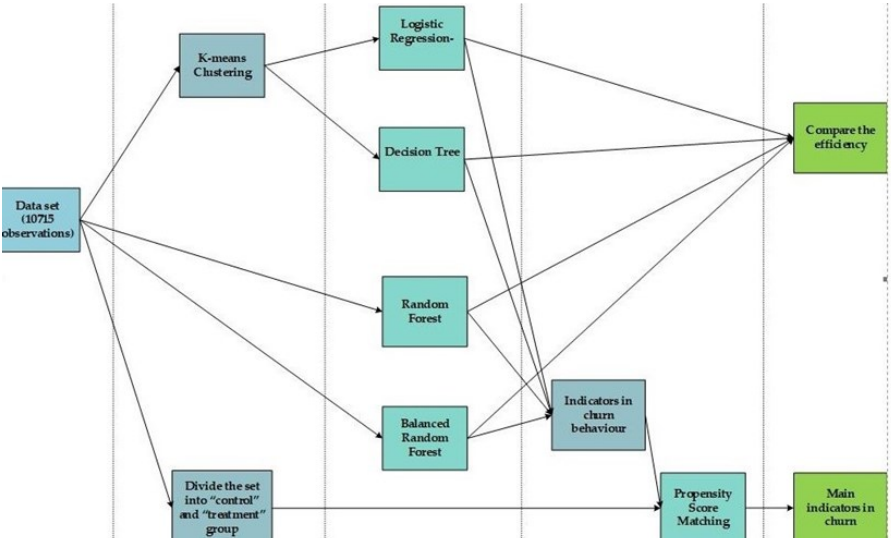

Data set (10715 observations)
K-means Clustering
Logistic Regression-
Decision Tree
Random Forest
Balanced Random Forest
Divide the set into "control" and "treatment" group
Indicators in churn behaviour
Propensity Score Matching
Compare the efficiency
Main indicators in churn

Figure 3. Analysis steps

In this paper, 5 data mining techniques are used, of which: 4 predictive modelling techniques (logistic regression, decision trees, random forest and balanced random forest) and 1 matching technique - used to estimate the net effect of a treatment (validation of the most important factors decision in churn action).

Regarding the methodology, the Logistic Regression techniques will be first used in churn predictions. In this case, the Logistic regression serves in the calculation of the precision rate, based on the ROC curve and the value of the AUC coefficient. Also, we are interested in determining which of the analyzed indicators has the greatest impact on the phenomenon studied and to what extent. The second situation is based on the use of the decision trees.

Last, Random Forest, Balanced Random Forest and Propensity Score Matching will be applied to the entire original data set. Random Forest and Balanced Random Forest serves to extract the main classification characteristics and calculate the importance of the nine variables on the churn action, while Propensity Score Matching will be applied to test the impact of a treatment that consists of a contact policy to customer instability, which has the effect of churning. On this purpose, 1505 individuals which have not previously been contacted by the telecoms service provider for 1 year have been selected in the control group.

Thus, due to the unbalanced distribution, the data set has been divided into groups as homogeneous as possible so that we can analyze the individual behavior, depending on the group it belongs to. The behavior of churn is studied within these classes obtained by the k-means clustering method. The grouping criterion was based on the average from invoice and the additional cost, grouped into three slightly disproportionate classes. The first group has been made up by the waste majority of the individuals, counting 8032 (74.96%), the second cluster is made up by 246 individuals (2.3%), while the third cluster contains 2437 individuals (representing 22.74% from the whole data set).

The churn rate for the first group is 13.07%, 1050 individuals, for the second one is 23.98%, 59 individuals, while for the third group the churn rate is of 14.73%, namely 359 individuals. It can easily be observed that the first cluster retains the highest number of individuals and presents a high lever of homogeneity when compare to the second and third cluster. On the other hand, the second cluster presents a high heterogeneity among the individuals, a high number of them having a behavior that can be associated to an "outlier" behavior. Based on these data it can be stated that the churn rate increases with the invoiced amount, while the additional cost is also rising. Therefore, in order to obtain high accuracy results, we will apply logistic regression and decision trees to the three homogeneous clusters resulting from k-means clustering.

3.3. Churn prediction and variables identification

In the context of prediction in the telecommunications industry, the phenomenon of churn is not a common event. In all sets of data studied and analyzed in specialized literature, the percent for non-churners customers is higher than that of churners clients.

3.3.1. Logistic regression

According to the literature, estimates that measure model performance should be rebalanced in unbalanced learning. The accuracy rate we will report is the ROC curve and, implicitly, the value of the AUC coefficient. Due to small sample of the churners compared to the non-churners, the whole dataset has been considered into just one group, rather than dividing it into the initiation set and test set.

In the prediction model, represented in this paper were included several variables, more precisely all the indicators were included, but only those with p value <0.05 were selected and displayed as output in the tables below.

Table 1. Results of logistic regression for first cluster

Coefficients	Estimate	Std. Error	Z value	PR(>|z|)	Exp(coef)
(Intercept)	-0.763	0.347	-2.198	0.027	0.466
Tenure	-0.009	0.002	-4.847	1.25E-06	0.99
MonthsO	0.012	0.001	10.753	<2e-16	1.012
MinR	0.537	0.162	3.319	0.0009	1.710

The characteristics that influence a customer's likelihood of churn are: the length of time the customer has used by the company's services, months number since changing his last offer (MonthsO), and minutes number consumed outside the company (MinR). Regarding the other variables, no important differences between the distribution of churners customers and non-churners customers has been identified. Also, the AUC is 0.604, indicating a low predictive capacity.

We followed the same procedure for the second cluster. The obtained results are presented in Table 2. Based on the received data, it has been observed that the invoice value paid for the services used as well as off-site minutes (Invoice) have shown significant importance for this segment of customers. The AUC for this model is 0.722.

Table 2. Results of logistic regression for the second cluster

Coefficients	Estimate	Std. Error	Z value	PR(>|z|)	Exp(coef)
(Intercept)	-1.087	1.045	-1.040	0.298	0.337
Invoice	-0.009	0.013	0.712	0.476	1.009
MinR	0.406	0.594	0.683	0.495	1.501
Tenure	-0.004	0.005	-0.807	0.420	0.995

For the third cluster, the results are presented in Table 3. Based on the data, it can be observed that in the case of the third cluster, the value of the extra cost paid for off-site services (ExtraCosts) has been of great importance to this segment of customers.

Table 3. The results of the logistic regression for the third cluster

Coefficients	Estimate	Std. Error	Z value	PR(>|z|)	Exp(coef)
(Intercept)	-1.137	0.497	-2.286	0.022	0.320
MonthsO	0.030	0.004	6.455	1.08E-10	1.031
Tenure	-0.007	0.002	-2.597	0.009	0.992
ExtraCosts	0.036	0.021	1.69	0.090	1.037

The AUC for the model is 0.742. The AUC results of the all three clusters are around 0.7, which means that the prediction of churn customers has an average accuracy and it can be improved.

3.3.2. Decision tree

The primary purpose of the algorithms underlying decision trees is to use a divisivity criterion to determine the most predictive factor and location as the decision point in the tree, and then to perform a predictive search to build the sub-tree until there is no data to be processed. Thus, by applying this data mining technique to the studied data set, we aim to identify which of the all the characteristics involved is the most predictive factor in the churn model, as well as the classification of the other indicators as regards the churn process. Basically, the classification question we are addressing in the case of this method is: What are the factors that influence churn action?

The decision tree of the first group, and from all the characteristics involved in the analysis, it identifies MonthsO and MinR var as being the most important in the churn prediction. Thus, it can be observed that among the two variables, the time measured in months number since the last account change (MonthsO) is the most predictive factor power , which can influences 88% of the churn for the first cluster. The second factor, represented by the received from other national company competing with the telecommunication provider (MinR), influences only 12% of the churn process – Table 4.

Table 4. The importance of variables. in the. model for .the first .cluster

Variable Importance	
MonthsO	MinR
88	12

The decision tree presented also highlights the number of months since the last offer (shift of service) in the churn action as the first decision point (MonthsO). Based on the obtained results it can be stated that a customer which has not been contacted by company to change actual offer and, practically, his last service has been renewed about 40 months ago will mostly prone to churn. In our case, of the 1472 individuals presenting this risk, 72.49% were classified correctly. As for the MinR variable, it has been shown using the decision trees that when a customer is near to fulfill a year in the network (more specifically, when he/she has 11.8 months of contract with the current telecommunication company), the risk of churn arrizes – a percentage of 72.97% of the persons having this characteristics have been correctly classified in this case. Nevertheless, it has been observed that if the number of minutes a customer uses for speaking with other persons outside his/her network surpasses 73.47% of the total numer he/she uses whin the network, he/she is very possible to act as a churner in the very near future. In this case, the value of the coefficient showing the model performance, AUC, for this cluster is 0.645, lower than the 0.7 imposed threshold.

With regard to the second tree generated, the value of the AUC coefficient exceeds the threshold of 0.7, reaching 0.796. In this case, the prediction put the accent on the second cluster and the results can be used in analysis and in the implementation of the measures needed to be taken for reducing or stopping the churn phenomenon.

The decision tree for this cluster present that the first decision maker in churn action is the same factor as in previous classes: the number of months since the subscription changed (MonthsO). The difference when compared to the first case is the multitude of predictive indicators that influence the phenomenon of churn – Table 5.

Table 5. Importance of variables to model for the second cluster

Variable Importance					
MonthsO	Invoice	MinR	ExtraCosts	Tenure	Age
29	22	21	21	5	2

Like in the model for estimating and classifying predictive factors for churn used for first cluster decision tree type method, in this case the churn pattern rule is given by the duration of the contract (MonthsO), the value of the bill (Invoice), the percentage of minutes received from other national networks (MinR) and the additional cost paid play (ExtraCosts).

Regarding the values obtained through the decision tree, it can be observed that in this case, the persons who haven't had a re-offer over their contract in the last 37.3 months are prone to churn. Even more, it has been observed that all the persons with an additional cost higher than 15.28 euros and a percentage of received calls of more than 82.06% are churners and the decision tree succeeds in classifying them correctly as churners. Also, it has been observed that if the invoice is higher than 51.21 euros, while the extra cost is between 15.28 euros and 37.3 euros, the individuals are susceptible to have a churn behavior and the decision tree classified them correctly with an accuracy higher than 65%. There are also some exceptions from the underlined rules as even some individuals having the percentage of MirR smaller than 19.78% and the additional costs (ExtraCosts) smaller than 6.75 euros, are also predisposed to churn.

The AUC of the third cluster model, 0.757, is a value very similar to that generated by the logistic regression.

Table 6. The importance of variables in model for cluster three

Variable Importance	
MonthsO	ExtraCosts
90	10

The decision tree for the third cluster identifies as the most important indicators in the prediction model of the MonthsO and ExtraCosts, with an importance coefficient of 90% for the first indicator, MonthsO, and 10% for the second, ExtraCosts. It has been determined that, for this group, if the MonthsO is higher than 31.74 months, the individuals are proned to churn, while the classification precision is of 80.26%.

3.3.3. Random Forest

In the following the Random Forest will be used for classifying the variables we have in the analysis by the degree to which they influence the churn process, thus observing which of these indicators "impose the rules" when a customer decided to churn. The Random Forest has been chosen as it is considered to bring better results than traditional classifiers, which would allow for a correct diagnosis with a very high degree of accuracy and which will help in automatically selecting the main features of the model, depending on their importance. Random Forest can also be applied to bootstrap samples, each developing a "non-trimming" tree, randomly selecting a certain number of predictors for each node, thus estimating the error, strength and correlation between the indicators analyzed.

The Random Forest's efficiency was tested on our data set. For this, in this paper was made the estimation using a traditional model. Therefore, the data set has been divided into a test set and a training set by randomly selecting 40% of the individuals for the first set (4286 individuals), while for the latter we allotted the remaining 60% (6429 individuals).

The target indicator in the model is the variable CHURN, and for each of the two levels of the vector we have the following two categories:

- on "Yes" - 1468 customers who have churned over the past few months;
- on the second level "No" - 9247 active clients who did not take part in this action.

Table 7. Random Forest Churn Classification Model

Call: randomForest formula = CHURN~ , data = train
Type of random forest : classification Number of trees : 500 No. of variables tries at each split : 3
OOB estimate of error rate : 14.15%

The pattern generated on the validation set is shown in Table 7, where it can be noticed that the type of method applied by Random Forest on the model is classification. Thus, it divides the dataset into 500 trees, classifying each of the three main features, with an estimation error of 14.15%, which means that the model has a prediction accuracy rate of about 85.85%.

Using the prediction function of the package we simulated a model prediction, which we compared with the current class of the CHURN vector on the validation set. The two entities contain the same values, which means that we have made the correct cassation, all six predictions are 100% correctly evaluated. We also notice from the matrix of confusion, which gives us a series of information about the model as the first class, that we have 6505 correctly classified individuals, and for the second class, we have 776 well-ranked churners from the validation set. The classification of the model has a high accuracy rate of 97.08%.

Table 8 presents a series of indicators determined as suggested by Amin et al. (2017). In our case the recall measures for the fraction of churners who are right identified like churners. The value of this indicator is 1 for the validation set, signifying that all the churners are identified as churners. Specific fraction measures for true non churners who are correctly marked as non-churners. The value of this indicator is high on the validation set as only 219 persons are misclassified. Precision measures the fraction of correct predicted churners over the total number of churn clients predicted by the model – also this indicator is high in the validation set, reaching 0.7799. The misclassification indicators refer to the instances where the persons of a class are classified in a class they do not belong to. Type-I Error is determined as the difference between 1 and Specificity, while Type-II Error is the difference between 1 and Sensitivity. Both indicators reach low values. Last, the F-Measure is a specific measure of precision and recall, determined as a weighted average of precision and recall (Amin et al., 2017). The recorded value in this case is 0.8763.

Table 8. The Confusion Matrix for validation set

Confusion Matrix and Statistics			
		Reference	
		Non-churn	Churn
Prediction	Non-churn	6505	219
	Churn	0	776
		Accuracy: 0.9708	
		Sensitivity (Recall): 1	
		Specificity: 0.9674	
		Precision: 0.7799	
		Misclassification: 0.0292	
		Type-I Error: 0.0326	
		Type-II Error: 0	
		F-Measure: 0.8763	

By applying the model on the test set it can be observed that the accuracy rate decreases to 85.19%. Particularly, the classification for the second class records larger errors, containing only 7 individuals correctly classified. The remainder of the indicators are presented in Table 9.

Table 9. The Confusion Matrix for test set

Confusion Matrix and Statistics

		Reference	
		Non-churn	
Prediction	Non-churn	2732	466
	Churn	10	7

Accuracy: 0.8519
Sensitivity (Recall): 0.4118
Specificity: 0.8543
Precision: 0.0148
Misclassification: 0.1481
Type-I Error: 0.1457
Type-II Error: 0.5882
F-Measure: 0.1671

Next, we aim to find out which of the data set variables impose the "rule" in the model. So, we used the varImpPlot function to find out what are the main features that influence churn action. This information is highlighted by the two averages generated: MeanDecreaseAccuracy – The Medium Accuracy Index, which tests how much the performance of the model will be affected without the default variables, and in our case the maximum importance in the accuracy of this prediction is the variable that expresses the minutes consumed by the client outside the network (MinC) and MeanDecreaseGini – the Gini Index, which measures how pure the nodes are at the end of the tree without each variable. Thus, based on the two indexes we can conclude that the highest influence in churn modeling is given by the minutes used for making phone calls in other national networks (outside the network the customer has contract with, (MinC) and the age indicators: both the age of the client and the lifetime of this contract with the telecommunication company (Age and Tenure).

3.3.4. Balannced Random Forest

In this case we have applied the Balanced Random Forest on the data set divided as before: 60% in the validation set and 40% in the test set. The model has been build dased on 1000 instantiated iterations, having as target variable the Churn variable.

Table 10 presents the results obtained for the test set. It can be observed that in this case, the values obtained for Recall is higher than in the previous case, signifying that the fraction of churn clients who are true recognize as real churn is higher in this case. Due to the fact that the set is characterized by heterogenous data, the obtained results are in line with our expectations.

Table 10. The Confusion Matrix for test set

Confusion Matrix and Statistics

		Reference	
		Non-churn	
Prediction	Non-churn	2234	
	Churn	115	266

Accuracy: 0.7776
Sensitivity (Recall): 0.8365
Specificity: 0.7883
Precision: 0.3072
Misclassification: 0.2224
Type-I Error: 0.2117
Type-II Error: 0.1635
F-Measure: 0.4494

Further on, we have considered all the variables and we have conducted a permutation importance analysis. The results are presented in Figure 4. It can be observed that from all the variables used in the churn model based on Balanced Random Forests, the most important variables listed are: MonthsO and MinC, those being also the characteristics that best discriminate between churners and non-churners.

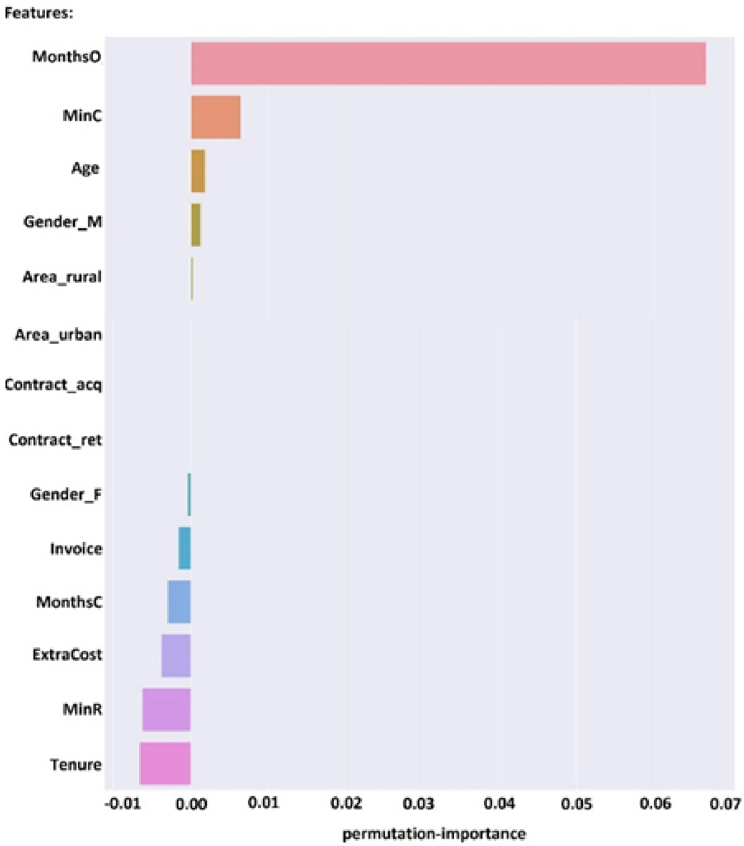


Figure 4. Features permutation importance

3.4. Propensity Score Matching

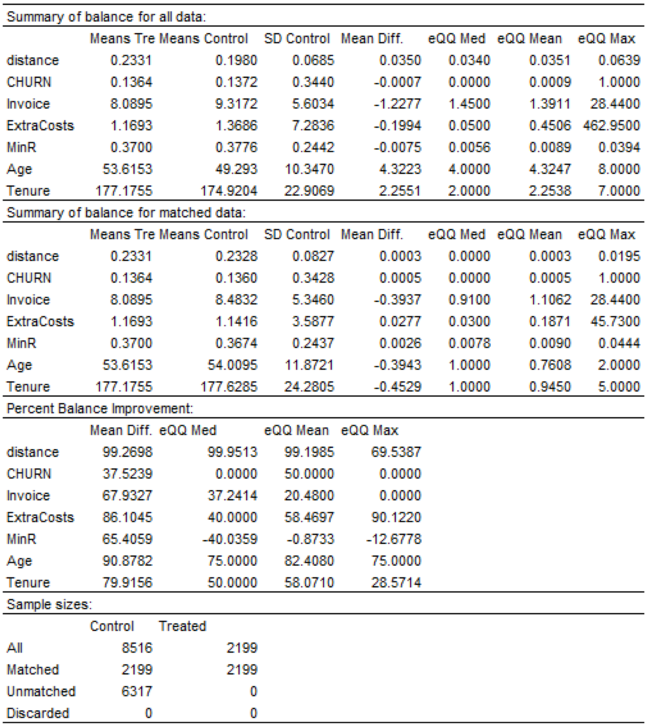
This analysis estimates the net effect on a result. It is used to compare two different groups (target group vs. control group). Propensity Score Matching (PSM) uses an estimated probability of group membership, e.g. control group based on observed predictors, usually obtained from logistic regression. Inclination scores can also be used for matching.

Through PSM we want to validate the main indicators obtained by modeling. This is strictly an original idea, not inspired by other works. We wanted to implement this methodology inspired by the policy applied in the telecommunications company: that 12 months contact policy, considered in our work as the treatment applied in the PSM. Above all, we want to demonstrate that MonthsO is a major deciding factor in churn action, as it has also been done by applying predictive modeling techniques.

On our dataset we apply the "treatment" called the 12-month contact policy. To this aim, we will select 1505 individuals in the control group that have not been contacted by the telecommunication service provider for 1 year. In this context, the "treatment" is made through the MonthsO variable. We applied this treatment on 14% of the data set, wanting to find out if it significantly influences the churn process.

In the analysis, we are particularly interested in the variables Churn, but also the Invoice and ExtraCosts, which represent the churn action, the customer invoice value and the additional cost that are influenced or not by the applied treatment. We will compare the individuals in the treatment group to those in the control group, but after matching them on the basis of the characteristics. A logistic regression has been estimated, in which the explanatory variable is the treatment variable. Based on the linear predictor, the distance between 2 individuals has been calculated. The nearest neighbor's method randomly moves from individual to individual, looking for the closest person in the control set and matches it. In Table 11 we can see how close the distributions (mean diff) are before and after this matching.

Table 11. PSM model - difference between treated and untreated



```
Summary of balance for all data:
           Means Tre Means Control SD Control Mean Diff.  eQQ Med eQQ Mean  eQQ Max
distance      0.2331        0.1980     0.0685     0.0350   0.0340   0.0351   0.0639
CHURN         0.1364        0.1372     0.3440    -0.0007   0.0000   0.0009   1.0000
Invoice       8.0895        9.3172     5.6034    -1.2277   1.4500   1.3911  28.4400
ExtraCosts    1.1693        1.3686     7.2836    -0.1994   0.0500   0.4506 462.9500
MinR          0.3700        0.3776     0.2442    -0.0075   0.0056   0.0089   0.0394
Age          53.6153        49.293    10.3470     4.3223   4.0000   4.3247   8.0000
Tenure      177.1755      174.9204    22.9069     2.2551   2.0000   2.2538   7.0000
Summary of balance for matched data:
           Means Tre Means Control SD Control Mean Diff.  eQQ Med eQQ Mean  eQQ Max
distance      0.2331        0.2328     0.0827     0.0003   0.0000   0.0003   0.0195
CHURN         0.1364        0.1360     0.3428     0.0005   0.0000   0.0005   1.0000
Invoice       8.0895        8.4832     5.3460    -0.3937   0.9100   1.1062  28.4400
ExtraCosts    1.1693        1.1416     3.5877     0.0277   0.0300   0.1871  45.7300
MinR          0.3700        0.3674     0.2437     0.0026   0.0078   0.0090   0.0444
Age          53.6153       54.0095    11.8721    -0.3943   1.0000   0.7608   2.0000
Tenure      177.1755      177.6285    24.2805    -0.4529   1.0000   0.9450   5.0000
Percent Balance Improvement:
           Mean Diff.  eQQ Med  eQQ Mean  eQQ Max
distance      99.2698   99.9513  99.1985  69.5387
CHURN         37.5239    0.0000  50.0000   0.0000
Invoice       67.9327   37.2414  20.4800   0.0000
ExtraCosts    86.1045   40.0000  58.4697  90.1220
MinR          65.4059  -40.0359  -0.8733 -12.6778
Age           90.8782   75.0000  82.4080  75.0000
Tenure        79.9156   50.0000  58.0710  28.5714
Sample sizes:
           Control  Treated
All           8516     2199
Matched       2199     2199
Unmatched     6317        0
Discarded        0        0
```

For example, the distance between treated and untreated (control group) was 0.0351 before the match and then decreased to 0.0003. It can also be observed that the age characteristic narrowed from 4.32 to -0.39.

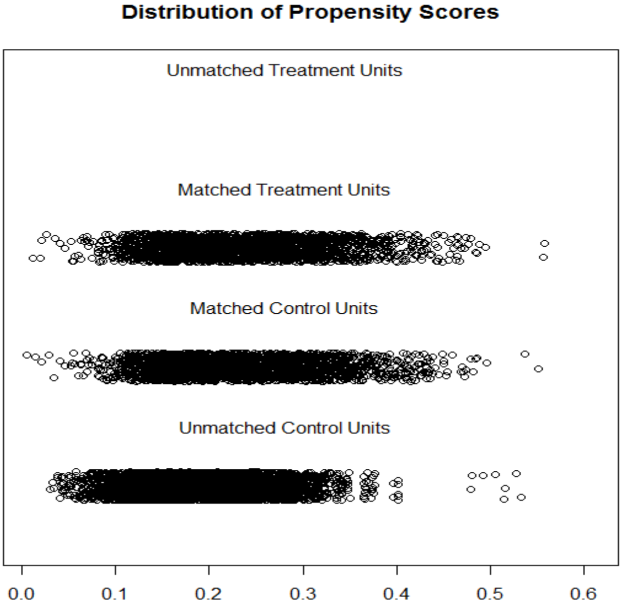


Figure 5. Distribution of propensity scores

Figure 5 highlights the fact that not all the individuals in the control group have been matched. Thus, out of a total of 10715 observations, 8516 have been assigned to the control group, representing 79.48% of the total, while the other 2199 have been exposed to the "treatment", meaning that it has been assumed that these individuals have not been contacted for a whole year by the telecommunications company. From the 8516 individuals in the control group, only 25.82% were matched, the remaining 6317 individuals remain unmatched, while of the 2199 treated individuals, all of them have been matched.

It can be noticed that the distributions are similar for treatment and control, which means that matching (combining the two groups) has been successful.

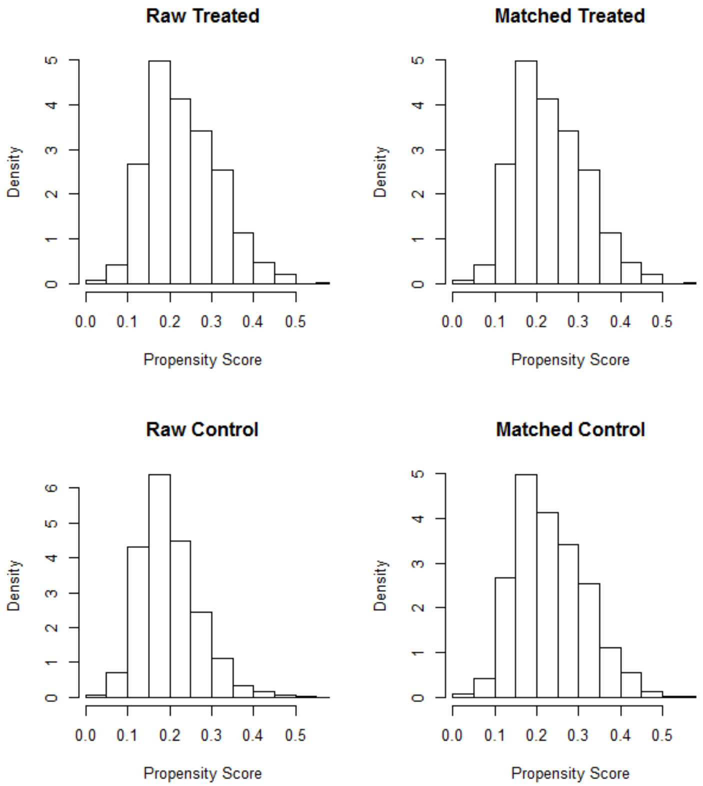


Figure 6. Distribution of scores before and after matching for treated and untreated objects

From the histograms in Figure 6, for individuals who have followed the treatment, that is, the individuals have not been contacted for 12 months by the telecommunication service provider, but also for the other individuals, both before and after matching, it can be observed that most of the values are in the range of [0.15; 0.2], and the lowest values are between [0.0; 0.1] or [0.45; 0.5]. It can also be said that distribution is approximately normal, with a downward trend, both for the individuals who have been treated and for the others, in both cases: before and after matching.

Table 12. PSM Model - Net Effect of Applied Treatment

Model					
Call: z5$zeling(formula = CHURN ~ Invoice + ExtraCosts + MinR + Tenure + MonthsO, data=match.data(rez1, "control"))					
Residuals					
	Min	1Q	Median	3Q	Max
	-0.54572	-0.14641	-0.11480	-0.07515	0.97456
Coefficients					
	Estimate	Std. Error	t value	Pr (>|t|)	
(Intercept)	0.2383	0.0554	4.301	1.78E-05	
Invoice	0.0037	0.0015	2.507	0.0122	
ExtraCosts	0.0005	0.0022	0.255	0.7984	
MinR	0.0428	0.0297	1.441	0.1498	
Tenure	-0.0011	0.0002	-3.858	0.0001	
MonthsO	0.0023	0.0002	8.850	<2e-16	
Residual standard error: 0.3359 on 2193 degrees of freedom					
Multiple R-squared: 0.04242					
Adjusted R-squared: 0.04024					
F-statistic: 19.43 on 5 and 2193 DF					
p-value: <2.2e-16					

Based on the data in Table 11 it can concluded that months number since the last change on their account (MonthsO), the average value of the invoice (Invoice) and network age of the customer, expressed in months (Tenure) are the three indicators explaining the net effect produced by the applied treatment (namely, the customers have not been contacted for 12 months) on the churn action, as the p-value for the three variables is lower than the maximum confidence level accepted, 0.05. A 12-month contact policy, the treatment that some of individuals have undergone, has affected the customers' decision to leave the network.

4. Discussion

4.1. Discussion about performance, accuracy, validation models

Logistic regression: Based on the analysed data it can be observed that the AUC of the first customer group is 0.604, indicating a low churn prediction capacity. For the second cluster the coefficient showing the performance of the model increased from 0.722. The AUC for the third group is bigger than the previos two clusters, 0.742, but very similar in value to group 2. In this context of telecommunication prediction is difficult because the data sample are not always balanced, the churn rate being much lower than the rate of active customers.

Decision Trees: In this case, the AUC of the first cluster is 0.645 and 0.696 of the second group. Regarding the third generated tree, which show the classification and estimation of predictive factors in the second cluster, the AUC value is above the threshold value of 0.7, reaching 0.757, which means that the model can be used for predictions.

Random Forest and Balanced Random Forest: In the case in which the Random Forest has been applied, the classification of the model on the test set has a high Accuracy rate of 85.19%, with a low Recall of 41.18%. However, the model based on the Balanced Random Forest has an acceptable Accuracy of 77.76%, with a high Recall of 83.65%. As the telecommunication companies are maily interested not to lose the churn customers, we will consider the Recall indicator for discriminating among the two methods, leading us to prefer the Balanced Random Forest for the particular seleced data set.

Thus, out of all four techniques applied on the dataset, Balanced Random Forest gave the best performance in predicting and ranking customers churn. This is demonstrated by the generation of a higher metric.

4.2. Discussions on reaching the set goals

Our original contribution, of the authors, but also of the work consists in the study and analysis of the Romanian data set in the field of telecommunications churn. The number of studies that are published in recent years about the prediction of churn in telecommunications companies proves that this problem has become a major one.. Another objective is to identify the main decision makers in the Romanian churn, but also to test the scientific techniques of the data consecrated in the literature regarding this aspect. I have another new aspect that the paper presents is the mix of methods implemented.

According to the logistic regression, the characteristics that affect a customer's churn probability are: the length of time the customer used the Tenure, the number of months since the change in his last offer (MOnthsO), the number of minutes received outside the company for first group (MinR). For the second and third clusters, the value of the invoice paid for the services used (Invoice), the minutes received outside the network (MinR), the value of the extra cost paid for off-network services (ExtraCosts) are of great importance to the phenomenon studied.

Decision trees release the variable MonthsO as the prediction factor with the highest impact in the churn predicting model, classifying it as the first decision in all three groups. From this we can conclude that the answer to the question: What are the features that influence the churn action? – is represented by the months number since the last offer was changed, followed national minutes received, the additional cost and the invoice value are the main features that influence the churn action.

Random Forest and Balanced Random Forrest pattern response at: Which of the data set variables requires the rule in the model? is expressed by the MinR variable in the case of the Random Forest and by the MonthsO and MinC in the case of Balanced Random Forest.

Propensity Score Matching determines the months number since the last change on actual account (MonthsO), the average invoice value (Invoice) and customer seniority (Tenure) as the three indicators explaining the net effect that the treatment applied, namely: customers have not been contacted for 12 months, on the churn action.

5. Conclusions

Telecoms are one of the sectors where the clients base represents a significant scope in maintaining stable revenues, with attention being paid to preventing their migration to other competitors. Its efficient management in the telecommunication industry is an extremely active field at global level, the churn action management initiative being taken to data mining and data scientists specialists in most Western countries.

The number of studies published in last years on the prediction of churn in telecommunications show that this issue has becomed and has continued to be a major concern over the years.

This paper addresses this widely discussed topic, the prediction of churn on a dataset of a major telecommunication company in Romania. The paper analyzes the behavior of the customers in a telecommunication company through five statistical methods of modeling and prediction, having as main objective the anticipation of the main variables that might give a signal regarding the individuals who might give up the services offered by the company analyzed for other telecommunication services from a competing company. In addition to analyzing this behavior (churn), the paper also focuses on the efficiency and performance of the data mining techniques in predicting churn behaviour in order to detect in a timely manner and with as high a degree of accuracy this action as the company can prevent the instability that is going to occur.

Subsequent research directions include the implementation of some analyzes and statistical models based on new prediction and management strategies for churn action. The main challenge being the need to identify the optimal, efficient and modern data science techniques to deal with this phenomenon and impacti. Depending on the method applied, low values for

AUC can be encountered, which makes it difficult to can make a accurately predict the churn phenomenon. However, the decision trees, the technique used on the calibrated data set, outlines the variable MonthsO as the predictive factor with the greatest importance in the churn predictiv model, classifying it as the first decision point, meaning that the months number since changing the last offer is the main indicator that influences churn action. After applying the Random Forest method, we can conclude that the greatest influence in shaping the churn is given by the minutes used in other national networks (MinC), while the Balanced Random Forest lists MonthsO and MinC as the leading variables in churners identification. PSM, the last applied technique determines the months number since the last change on the actual account (MonthsO), the average value of the invoice (Invoice) and seniority of the client (Tenure) as the three indicators explaining the net effect that the applied treatment has on the churn action. Based in these results, we aim in the future to extend the data set and to include the customers having pre-paid Sims.

Therefore, the most important decision factor in the churn action of the client in the telecommunications sector in Romania is the number of months since the last offer, included in the variable MonthsO, a variable that appears as the first decision factor in all the tested techniques.

References

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Alawfi, K., Hussain, A., & Huang, K. (2017). Customer churn prediction in the telecommunication sector using a rough set approach. *Neurocomputing*, *237*, 242-254. https://doi.org/10.1016/j.neucom.2016.12.009

De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*, *269*, 760-772. https://doi.org/10.1016/j.ejor.2018.02.009

Euler, T. (2006). Churn prediction in telecommunications using miningmart. *Proceedings of the workshop on data mining and business (DMBiz)*, 1-2. https://sfb876.tu-dortmund.de/PublicPublicationFiles/euler_2005c.pdf

Faris, H. (2018). A Hybrid Swarm Intelligent Neural Network Model for Customer Churn Prediction and Identifying the Influencing Factors. *Information*, *9*(11), 288. http://dx.doi.org/10.3390/info9110288

Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. *Expert Systems with Applications*, *39*(1), 1414-1425. https://doi.org/10.1016/j.eswa.2011.08.024

Hung, S.-Y., Yen, D.C., & Wang, H.-Y. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications*, *31*(3), 515-524. https://doi.org/10.1016/j.eswa.2005.09.080

Kisioglu, P., & Topcu, Y. I. (2011). Applying Bayesian Belief Network approach to customer churn analysis: A case study on the telecom industry of Turkey. *Expert Systems with Applications*, *38*(6), 7151-7157. http://dx.doi.org/10.1016/j.eswa.2010.12.045

Kotler, P., & Keller, K.L. (2016). *Marketing management* (15th ed). Pearson.

Mattison, R. (2005). *The telco churn management handbook*. Xit Press.

Mozer, M. C., Wolniewicz, R., Grimes, D. B., Johnson, E., & Kaushansky, H. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Transactions on Neural Networks*, *11*, 690-696. https://doi.org/10.1109/72.846740

National Authority for Administration and Regulation in Communications Ancom. (2016). *Annual report 2016.* https://www.ancom.ro/en/uploads/links_files/20170410_Raport_Anual_2016_integral_en_FINAL.pdf

Qi, J., Zhang, Y., Zhang, Y., & Shi, S. (2006). TreeLogit Model for Customer Churn Prediction. In C.-J. Tan & L. Zhang (Eds.), *Proceedings of the 2006 IEEE Asia-Pacific Conference on Services Computing (APSCC'06)* (pp. 70-75). IEEE. https://doi.org/10.1109/APSCC.2006.111

Tsai, C.-F., & Lu, Y.-H. (2009). Customer churn prediction by hybrid neural networks. *Expert Systems with Applications*, *36*(10), 12547-12553. https://doi.org/10.1016/j.eswa.2009.05.032

Wei, C.-P., & Chiu, I.-T. (2002). Turning telecommunications call details to churn prediction: a data mining approach. *Expert Systems with Applications*, *23*(2), 103-112. http://dx.doi.org/10.1016/S0957-4174(02)00030-1

Zhao, Y., Li, B., Li, X., Liu, W., & Ren, S. (2005). Customer Churn Prediction Using Improved One-Class Support Vector Machine. In X. Li, S. Wang & Z. Y. Dong (Eds.), *Advanced Data Mining and Applications, Vol. 3584* (pp. 300-306). Springer. https://doi.org/10.1007/11527503_36