IJEEC

International Journal of Electrical Engineering and Computing
Vol. 2, No. 1 (2018)

Original research paper
UDC **339.13:621.39**
DOI **10.7251/IJEEC1801029M**

Performance comparison of six Data mining models for soft churn customer prediction in Telecom

Marin Mandić [1], Goran Kraljević [2], Ivan Boban [3]

[1] Faculty of Mechanical Engineering, Computing and Electrical Engineering, University of Mostar, Mostar, Bosnia and Herzegovina

marin.mandic@hteronet.ba, goran.kraljevic@hteronet.ba, ivan.boban1@hteronet.ba

Abstract— **Due to a high competition in the market, the telecom operators are affected by churn, therefore it is very important for them to identify which users are likely to leave them and switch to the competition telecom company. This research uses data on behaviour of the users from telecom systems that serve to identify patterns in behaviours and thereby recognize the churn. Creating new definition of prepaid soft churn based on multiple conditions is valuable contribution of this paper. At preparing data, a selection of useful attributes was made using the Principal Component Analysis (PCA). The normalization of the attribute values has also been made in order to obtain a proper balance of the influence of all the attributes. Common problem with telecom churn prediction data is imbalance, taking into account the target variable. Such a case is also in the data used in this paper, where the percentage of churners is 12%. Comparison of undersampling and oversampling was performed as a method for resolving the data imbalance problem. Data sets with undersampling and oversasmpling have been used to train the decision tree, logistic regression and neural network algorithms and therefore six prediction models for detecting the churn of the Prepaid users in the telecom were created in this paper. Performance analysis and comparison of the six developed Data mining models was also performed.**

Keywords- predictive modelling, Prepaid, Data mining, machine learning algorithms, churn

I. INTRODUCTION

Strong competition in the telecom market causes users' churn, where such users cancel to use services from one telecom operator and switch to a competition telecom. The churn causes revenue loss for the telecom, and therefore it is very important for the telecom to influence on the reduction of such churns. In order to prevent the churn, it is necessary to identify, with some probability, which users will switch to a competition telecom operator. The prediction process uses recognition of patterns in the users' behaviour to predict their future behaviour. Telecom companies have history data on their customers' behaviour and can use them in the churn prediction process. If a telecom manages to predict which users will leave it, then it can try to keep them with certain marketing campaigns. Marketing campaigns are activities focused towards the user in order to increase the user satisfaction. In this way, it is possible to keep existing users, which will lead to increase, i.e. retain of the revenues and profits for the telecom.

In their work, V. Lazarov and M. Capota [1] state that it is five to six times more expensive to bring a new user than to keep the existing one. This is another reason why it is important to keep users and identify those with the potential of churn to a competition telecom.

This paper is a revised and expanded version of the paper presented at the XVII International Symposium INFOTEH-JAHORINA 2018 [12].

Corespondence to: M. Mandic (marin.mandic@hteronet.ba)

The key contributions of this paper are:

- Defining soft churn in a unique way with multiple conditions that includes this large majority of real churn users
- Using real telco set of data and developing model that combines attribute selection and normalization with resolving class imbalance problem that predicts prepaid soft churn with high accuracy and sensitivity.

The rest of the paper is organized in such a way that Chapter II describes the current researches in the specific field, and Chapter III presents the created model for churn prediction, churn definition, experiment and the experiment results. Finally, Chapter IV presents the conclusion of the work and Chapter V presents the directions for further research.

II. CURRENT RESEARCHES

A lot of researches have been done in the field of churn prediction in the telecom industry.

In their work, Kiran Dahiya and Surbhi Bhatia [2] have defined five steps of the churn prediction process, namely:

a) data collection,
b) data preparation,
c) data pre-processing,
d) extraction of variables,

e) *implementation of machine learning algorithm*.

A precisely defined process helps in solving the complex problem of churn prediction. J. Zhang, J. Fu and other authors have defined a model for the early churn prediction in their work [9]. The model for early churn prediction enables more time for the telecoms to conduct an effective retention marketing campaign. It has been proven that churn prediction significantly decreases with increase of the time interval of prediction. In their work they have created a hybrid model for early churn prediction based on UTS, PCA MTS and Social Influence model, which has shown that it is possible to predict potential churn even several months in advance.

The group of authors in their work [3] has created a comparison of oversampling methods with the application to the problem of churn prediction. They compared the MTDF, SMOTE, ADASYN, TRkNN, MWMOTE and COTE methods. The results of their research have shown that it is best to use the MTDF method. This method helps in solving the problem of imbalance in the data, which is common in the churn prediction in a telecom. The problem of imbalance is that we have a significantly higher number of non-churners in relation to churners.

In their work, Hui Li, Deliang and other authors [4] have described the importance of selection of the attributes for creating the prediction model. The attribute selection enhances the model in three ways: simplifies the model and therefore makes it easier to interpret, shortens the model training time, and improves generalisation and thus reduces the excessive adaptability of the model to the training data. In their work, they used a random forest to measure the importance of a particular variable.

It is possible to use different methods for creation of a successful churn prediction model. The most commonly used methods are decision trees [2,4], logistic regression [2,5] and neural network [5,6]. In addition to the above-mentioned methods, Bayes Network [7], Support vector machines (SVM) [8] and various hybrid models [9] are also successfully used. The success of these models depends to a large extent on the data set on which the model is trained.

Backiel.Y. Verbinnen, B. Baesens and G. Claeskens in their work [10] applied the social network as additional information that increases the accuracy of the churn prediction. The analysis of the social networks is based on the social influence of some person on others. A person who is a churn has an influence on people from their social circle. They have proven that social networking data improve the accuracy by comparing two models, one that uses social networking data and a model that does not use social networking data.

III. CREATING THE MODEL FOR CHURN PREDICTION

A. *Input data*

Telecoms possess a lot of information about their users because they collect data about their behaviour. A large amount of these data can be redundant and cause lower accuracy of the model. In addition, a large amount of data causes the slowness of model execution.

The data used in this paper can be divided into:

1. User behaviour data that include attributes such as call duration within the network, duration of calls to other mobile networks, number of sent SMSs, amount of data transferred, and 50 other data related to user behaviour.
2. Data about the users themselves, not many of which are present in this paper, because these are prepaid (anonymous) users, therefore we do not have data such as gender, age, marital status etc. The attribute of the user's loyalty is used for this paper, that is, the time since the user have used the telecom services and the number of days until the expiration of the 'active user' status.

The data used for the experiment are shown in Table I.

TABLE I. INPUT DATA FOR THE EXPERIMENT

INPUT DATA	
Number of attributes	109
Number of records/users	49.868

The number of users indicated as churners is 6 328, which means there are 12.69% of the churners.

The common problem with prediction data is imbalance, taking into account the target variable. The group of authors has described this problem of imbalance in their work [3]. Such a case is also in the data used for this experiment, where the ratio of churners in relation to non-churners is 88:12.

The target attribute in the data set is defined as the binary type of data, so the churners are labelled as "true" and the non-churners as "false". In order for the models to have satisfactory accuracy, it is necessary to solve the problem of data imbalance.

In this experiment, a comparison of undersampling and oversampling was performed as a method for resolving the data imbalance problem, taking into account the target variable. Created data sets with undersampling and oversasmpling have been used to train the decision tree algorithm, logistic regression and neural network algorithms, and therefore six different comparison models have been obtained in this way.

Undersampling was done by undersampling of the non-churners, and 6 750 records were selected by random selection from the base of the non-churners.

70% of the records are used for model learning, and 30% of the records are left for testing.

70% of the records for learning contain 34.908 records. From that number of records there are 4.430 records labelled as the churners. All the churners are used for the experiment.

30.478 records were labelled as non-churners and 6.750 records were selected by random undersampling.

In this way, the set of the data for learning has obtained a ratio of churners and non-churners of 40:60. The goal of undersampling is to increase the accuracy of churn prediction.

At oversampling, we would like to obtain the same ratio of churners and non-churners of 40:60 as we did with undersampling. Since there are 34.908 examples in the training set labelled as non-churners, by using stratified oversampling of 4.430 examples labelled as churners, we would like to generate additional examples and obtain a total of 23.272 examples labelled as churners. The stratified random oversampling generates an additional set of examples, taking into account the class distribution so that the distribution is ultimately the same as in the entire set of examples. Such a training set of the total of 58.180 examples will be used to train the decision tree model, neural network and logistic regression models.

B. Churn definition and experiment

Defining postpaid churn is easy because postpaid churn can be voluntary churn, when a user cancels his subscription to current service provider and joins other service provider, and involuntary churn, when a user is cancelled by service provider due to unpaid bills or fraud. Voluntary churn can be divided into incidental and deliberate churn [13]. Former research paper mostly deals with defining postpaid churn where postpaid users must formally cancel subscription. A. Backiel, B.Baesens and G.Claeskens in their work [14] define a prepaid customer churn when customer have not made or received a call for more than 30 days. In our paper we define prepaid soft churn based on analysis of history data of prepaid users who permanently stopped using the telco services.

Prepaid churn is not easy to define because prepaid users cannot cancel contract obligations. We can explain prepaid churn issues by explaining the possible prepaid state.

Prepaid users can have multiple states:

- Active state when user can use and receive telco services (voice call, SMS, data usage etc.)
- Suspended state when a user has no money on prepaid account or account date is expired. Users in this state can only receive telco services
- Blocked state when a user cannot use or receive telco services. Blocked users are users who were in a suspended state and did not recharge his account for defined period of time.
- Canceled state have users who ware in blocked state did not recharge his account for defined period of time.

Users who are in suspended and blocked state can change their state to active by recharging their account.

It takes a lot of time for a prepaid user to get into a cancelled state, and in reality, this user has long stopped using that telco service. Therefore, it is necessary to set a definition of churn that will identify potential prepaid user churn as soon as possible, which is so-called a soft churn.

In our research, we define prepaid soft churn so we can identify as much prepaid users who are going to stop using service and those prepaid users who significantly reduced using services of telco service provider. We analyzed past period of data and conclude that we must implement multiple conditions to include the large majority of real churn users and not to include users that are still using telco services.

Our definition of prepaid soft churn includes one of the following three conditions:

a) The consumption on SMS and Voice services is less than 1 KM (KM is currency in Bosnia and Hercegovina),

b) the number of days of the activity is lower than the average for the last 3 months,

c) the average number of prepaid recharge for the last three months is less than 5 KM and there were no recharge in the month that defines the churn.

With this soft prepaid churn definition, we come to approximately 12% of churners. This percentage is pretty high, but relevant because prepaid user has no contractual obligation, so they can leave pretty easy. Prepaid marketing retention campaign cost are significantly lower than postpaid marketing retention campaign cost, so there is no much concern of this pretty high percentage of prepaid soft churn.

The input data set has 109 different attributes. Due to the large number of attributes, a selection of useful attributes was made, while the other attributes are not used to create a data mining model. As we noted in Chapter II Hui Li, Deliang and other authors [4] have used random forest to reduce the number of attributes and simplify the model. In our experiment we made selection of attributes by using the Principal Component Analysis – PCA. The Principal component analysis finds the direction in which we have the largest variance in a high-dimensional data set and is projected to a smaller set of data that retains most of the essential information from the initial data set. In this paper a selection of all the attributes with a weighting factor larger than 0.03 was made, so in this way selection has reached the final figure of 45 attributes for creating the Data mining model, which is shown in Table II.

TABLE II. LIST OF ATRIBUTTES

LIST OF ATRIBUTTES (after selection)
Conversation_Duration_For_Outgoing_Calls_To_VAS
Duration_Of_Incoming_Calls_From_Eronet
Duration_Of_Incoming_Calls_From_FBHT
Duration_Of_Incoming_Calls_From_M064
Duration_Of_Incoming_Calls_From_MTS
Duration_Of_Outgoing_Calls_To_Eronet
Duration_Of_Outgoing_Calls_To_Eronet_CallCenter
Duration_Of_Outgoing_Calls_To_FBHT
Duration_Of_Outgoing_Calls_To_FHT
Duration_Of_Outgoing_Calls_To_FTS
Duration_Of_Outgoing_Calls_To_M064
Duration_Of_Outgoing_Calls_To_MBHT
Duration_Of_Outgoing_Calls_To_MTS
Duration_Of_Outgoing_Calls_To_Other_BH_Mobile_CallCenter
Number_Of_Data_Transfered_Units
Number_Of_Distinct_Incoming_Eronet_Users
Number_Of_Distinct_Incoming_M064_Users

Number_Of_Distinct_Incoming_MBHT_Users
Number_Of_Distinct_Incoming_MTS_Users
Number_Of_Distinct_Outgoing_Eronet_Users
Number_Of_Distinct_Outgoing_FBHT_Users
Number_Of_Distinct_Outgoing_FHT_Users
Number_Of_Distinct_Outgoing_FTS_Users
Number_Of_Distinct_Outgoing_M064_Users
Number_Of_Distinct_Outgoing_MBHT_Users
Number_Of_Distinct_Outgoing_MTS_Users
Number_Of_Distinct_Outgoing_Other_Mobile_Users
Number_Of_Incoming_Calls_From_Eronet
Number_Of_Incoming_Calls_From_M064
Number_Of_Incoming_Calls_From_MBHT
Number_Of_Incoming_Calls_From_MTS
Number_Of_Outgoing_Calls_To_Eronet
Number_Of_Outgoing_Calls_To_Eronet_CallCenter
Number_Of_Outgoing_Calls_To_FBHT
Number_Of_Outgoing_Calls_To_FHT
Number_Of_Outgoing_Calls_To_FTS
Number_Of_Outgoing_Calls_To_M064
Number_Of_Outgoing_Calls_To_MBHT
Number_Of_Outgoing_Calls_To_MTS
Number_Of_Outgoing_Calls_To_Other_Mobile
Number_Of_Sessions
Number_Of_SMS_To_Eronet
Number_Of_SMS_To_M064
Number_Of_SMS_To_MBHT
Number_Of_SMS_To_MTS

The input data set contains the attributes that have significantly higher values than other attributes, for example, the "quantity of data transferred" attribute has higher values than the "number of calls" attribute. The attribute values depend on their nature and on the measurement units. The attributes with higher values have a greater impact on creation of the model. In order to make a proper balance of the influence of all the attributes, normalization has been done. By normalization, the attribute values rescaled within a certain range of numbers. Z transformation or, as it is also termed, statistical normalization has been done for the purposes of this paper.

The formula of statistical normalization is represented by the expression:

$$Z = (X - \mu) / s \tag{1}$$

Where:

- **X** - vector value of the attribute for which Z transformation is calculated;
- **μ** - the arithmetic mean that is subtracted from vector X;
- **s** - standard deviation

Z transformation converts data into normal distribution with mean value 0 and variance 1. In this way, all attribute values are reduced to the same scale and in this way it is possible to make a good comparison between them.

A comparison of several machine learning algorithms was done in the experiment.

All algorithms were trained and tested on the undersampling and oversampling data set and with a reduced set of attributes.

The algorithms used are: decision tree with algorithm C4.5, logistic regression and neural network.

It is necessary to note that decision tree is pruned to avoid overfitting. In our experiment we prepruned decision tree based on maximal depth of 26 node. Also pruning is performed after tree generation based on confidence level of pessimistic error calculation of pruning.

C. Experiment results

Standard measures have been used in this paper to compare the efficiency of different churn prediction models. The measures used in the paper are: accuracy, precision and sensitivity or recall.

With the markings hereinafter, we denote the examples in the confusion matrix and use them to calculate the standard model comparison measures:

a) TP - means accurately classified churners,

b) TN - means accurately classified non-churners,

c) FP - means inaccurately classified churners,

d) FN - means inaccurately classified non-churners.

$$\text{Accuracy} = (TP + TN) / (TP + FP + TN + FN) \tag{2}$$

The accuracy is a measure that is defined as the total number of precisely classified examples.

$$\text{Precision} = TP / (TP+FP) \tag{3}$$

The precision is the number of precisely classified churners in relation to the total number of examples labelled as churners. The higher the precision measure the smaller is the number of inaccurately classified churners.

$$\text{Sensitivity} = TP / (TP + FN) \tag{4}$$

The sensitivity or recall is the number of precisely classified churners in relation to the number of examples that really belong to the churner class. A higher value of the sensitivity indicates smaller number of actual churners that are inaccurately classified.

The confusion matrix presents the results of the experiment in the tabular form:

TABLE III. CONFUSION MATRIX

Class	actual non-churners	actual churners
predicted non-churners	TN	FN
predicted churners	FP	TP

The confusion matrix presents the results of the experiment, taking into account three algorithms: decision tree, logistic regression and neural network. Likewise, a comparison was done to address the problem of imbalance in the data by using undersampling and oversampling for each of the aforementioned algorithms. Graphic comparison of the algorithms used was also done with the ROC curve overview.

Table IV presents the confusion matrix for the model of the decision tree with undersampling:

TABLE IV. DECISION TREE C4.5 WITH UNDERSAMPLING

	FALSE	TRUE
FALSE	12403	316
TRUE	659	1582

The total accuracy of the decision tree model with undersampling is 93.69%, the sensitivity is 83.40% and the precision is 71.56%.

The decision tree model with undersampling presents the rules used for dividing the data into more homogeneous sets with respect to the targeted churn variable.

With the analysis of the constructed decision tree, it is possible to conclude that the most important attributes are the number and duration of calls within the network.

A gain ratio is used as the criterion for selection of the attributes for data partition.

Table V presents the confusion matrix for the model of logistic regression with undersampling:

TABLE V. LOGISTIC REGRESSION WITH UNDERSAMPLING

	FALSE	TRUE
FALSE	11962	439
TRUE	1100	1459

The logistic regression with undersampling has shown the worst results in predicting the churn. The total accuracy of the logistic regression is 89.71%, the sensitivity is 76.87%, and the precision is 57.01%.

Table VI presents the confusion matrix for the model of the neural network with undersampling:

TABLE VI. NEURAL NETWORK WITH UNDERSAMPLING

	FALSE	TRUE
FALSE	12098	286
TRUE	964	1612

The created neural network model with undersampling has one input, one hidden and one output layer.

The total accuracy of the neural network model with undersampling is 91.64%.

Neural network with undersampling has the highest value of the sensitivity rate of 84.93% i.e. of all the tested models it has the highest percentage of churn prediction. The neural network with undersampling has somewhat poorer precision rate of 62.58% compared with the decision tree.

Table VII presents the confusion matrix for the model of the decision tree with oversampling:

TABLE VII. DECISION TREE C4.5 WITH OVERSAMPLING

	FALSE	TRUE
FALSE	13016	1142
TRUE	46	756

The total accuracy of the decision tree model with oversampling is 92.06%, the sensitivity is 39.83%, and the precision is 94.26%.

The decision tree model with oversampling has shown worse results than the decision tree with undersampling, especially if we observe the sensitivity measure.

TABLE VIII. LOGISTIC REGRESSION WITH OVERRSAMPLING

	FALSE	TRUE
FALSE	12882	658
TRUE	180	1240

The created model of logistic regression with oversampling (Table VIII) has shown significantly better results in the churn prediction compared with the logistic regression with undersampling.

The total accuracy of the logistic regression is 94.40%, the sensitivity is 65.33%, and the precision is 87.32%.

TABLE IX. NEURAL NETWORK WITH OVERSAMPLING

	FALSE	TRUE
FALSE	12975	908
TRUE	87	990

The neural network model with oversampling (Table IX) has a slightly better total accuracy than the neural network model with undersampling, but has a significantly smaller sensitivity measure.

The total accuracy of the neural network with oversampling is 93.35%, the sensitivity is 52.16%, and the precision is 91.92%.

TABLE X. COMPARISION MACHINE LEARNING ALGORITAMS USED WITH UNDERSAMPLING AND OVERSAMPLING

Model	Acuracy	Sensitivity	Precision
Decision tree model with undersampling	93,69	83,4	71,56
Logistic regression with undersampling	89,71	76,87	57,01
Neural network with undersampling	91,64	**84,93**	62,58
Decision tree model with oversampling	92,06	39,83	**94,26**
Logistic regression with oversampling	**94,4**	65,33	87,32
Neural network with oversampling	93,35	52,16	91,92

From the results shown in the Table X, it is clear that the logistic regression with oversampling has the highest accuracy. The decision tree with oversampling has the highest precision, while the neural network with undersampling has the highest sensitivity value. It can be noticed from the Table X that the decision tree with undersampling has a high percentage of accuracy of all the measures, although it does not have the highest value of any of the calculated measures.

Figure 1 shows the relationship between the measures of accuracy, precision and sensitivity of each of the three machine learning algorithms used.

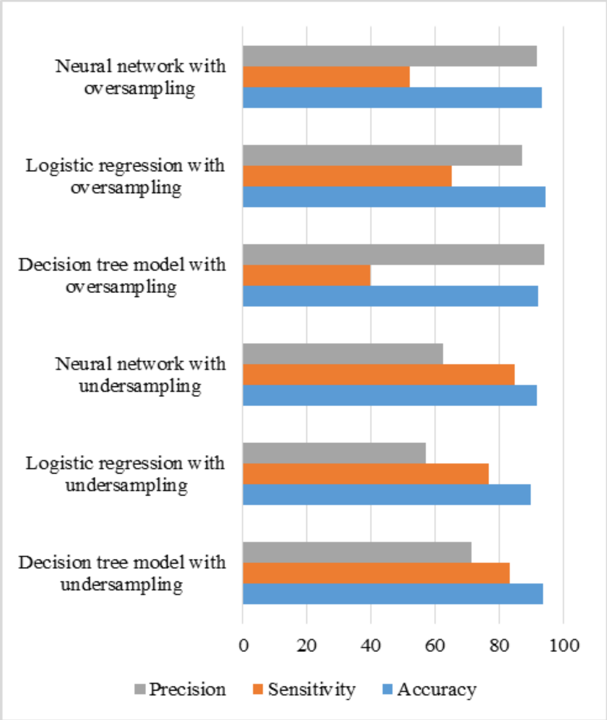

Figure 1 Display measures of the three machine learning algorithms used with undersampling and oversampling

Figure 2 shows the ROC (Receiver Operating Characteristics) curve which graphically shows the relationship between sensitivity and specificity, i.e. shows the relation of properly classified churners (TP) with respect to inaccurately classified churners (FP). A ROC curve was created for each of the three used machine learning algorithms with undersampling and oversampling.

Figure 2 ROC curve of the three machine learning algorithms used with undersampling and oversampling

The ROC curve algorithms are shown in the following colors:

- Decision tree model with undersampling - light green color;
- Logistic regression with undersampling - yellow color;
- Neural network with undersampling - light blue color;
- Decision tree model with oversampling - dark blue color;
- Logistic regression with oversampling - dark green color;
- Neural network with oversampling - red color;

The closer the ROC curve of the algorithm is to the point (0.1) the algorithm is better [11]. The created ROC curve shows that the decision tree with oversampling shows the best results, i.e. the best ratio of properly classified churners with respect to incorrectly classified churners. The ROC curve does not show that the decision tree with oversampling has a low percentage of predicted actual churners compared to the total number of actual churners, which makes this algorithm impossible to declare as the best.

The results of the experiment have shown that developed models of the decision tree, logistic regression, and neural network with undersampling generally have higher sensitivity values compared to developed models with oversampling. On the other hand, developed models with oversampling have significantly higher precision values. Given the problem of churn prediction that this experiment solves, it can be noticed that the decision tree with undersampling has high values of accuracy, sensitivity and precision measures and as such is the best constructed model.

IV. CONCLUSION

This paper presents a brief overview of recent research in this field over the last few years, also this paper is a revised and expanded version of the paper [12].

In this paper, a comparison of three different machine learning algorithms has been made in order to predict the prepaid telecom customers churn. The data set used has an imbalance in the data, observing the targeted churn attribute.

The emphasis of this paper is a unique prepaid soft churn definition, and developing a model that combines attribute selection and normalization with solving class imbalance problem with undersampling and oversampling. The statistical method Principal Component Analysis was used in this paper, which helped in reducing the number of attributes from 109 to 45.

A comparison of the performances of the models was made using the accuracy, sensitivity and precision measures. A confusion matrix and a ROC curve were also presented for each model.

The experiment showed that the decision tree with undersampling proved to be the best machine learning algorithm for this data set, because it has well-balanced values of the total accuracy, sensitivity and precision measures. The decision tree with undersampling has a high percentage of 83.4% of the churn guessing from the total number of churners, and at the same time there is not a large number of incorrectly classified churners.

Every step in the churn prediction process is very important, with particular emphasis on the steps of reducing the attributes and solving the problem of imbalance in the data.

V. FUTURE RESEARCH

Future research can be compared with additional machine learning algorithms for the given data set. It is also possible to examine the hybrid models of solving the problems of imbalance in the data in order to avoid the problem of losing information in the case of undersampling and the overfitting problem at oversampling. Since the churn prediction models are highly dependent on data, it would be good to use additional data sets for performance analysis of different models in future research. This research was used to test the quality of constructed models with a test data set that was not used for creating the model. Future research could test the constructed models with user behaviour data from the future period. In this way, the stability of the constructed models could be verified with a time lag.

LITERATURE

[1] V. Lazarov and M. Capota, “Churn Prediction,” Eighth ACM SIGKDD Int. Conf., 2007.

[2] K. Dahiya, “Customer Churn Analysis in Telecom Industry,” 2015.

[3] A. Amin et al., “Comparing oversampling techniques to handle the class imbalance problem : A customer churn prediction case study,” vol. 3536, no. c, pp. 1–18, 2016.

[4] H. Li, D. Yang, L. Yang, Y. Lu, and X. Lin, “Supervised Massive Data Analysis for Telecommunication Customer Churn Prediction,” pp. 163–169, 2016.

[5] S. A. Qureshi, A. S. Rehman, A. M. Qamar, and A. Kamal, “Telecommunication Subscribers ’ Churn Prediction Model Using Machine Learning,” pp. 131–136, 2013.

[6] V. Umayaparvathi and K. Iyakutti, “Applications of Data Mining Techniques in Telecom Churn Prediction,” Int. J. Comput. Appl., vol. 42, no. 20, pp. 5–9, 2012.

[7] C. Kirui, L. Hong, W. Cheruiyot, and H. Kirui, “Predicting Customer Churn in Mobile Telephony Industry Using Probabilistic Classifiers in Data Mining,” vol. 10, no. 2, pp. 165–172, 2013.

[8] L. Peng and Y. Xiaoyang, “Telecom Customer Chum Prediction Based on Imbalanced Data Re-sampling Method,” pp. 229–233, 2013.

[9] J. Zhang, J. Fu, C. Zhang, X. Ke, and Z. Hu, “Not Too Late to Identify Potential Churners : Early Churn Prediction in Telecommunication Industry,” pp. 194–199, 2016.

[10] A. Backiel, Y. Verbinnen, B. Baesens, and G. Claeskens, “Combining Local and Social Network Classifiers to Improve Churn Prediction,” Proc. 2015 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min. 2015 - ASONAM ’15, pp. 651–658, 2015.

[11] D. DeLong, and D. Clarke-Pearson, “Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach,“ International Biometric Society, 44(3):837–845, 1988.

[12] M. Mandic, G. Kraljevic and I. Boban, “Development of the Data mining model for churn prediction of Prepaid users in the telecom,” 17th International Symposium INFOTEH-JAHORINA, pp. 293–297, 2018.

[13] G.Kraljević, S. Gotovac, “Modeling Data Mining Applications for Prediction of Prepaid Churn in Telecommunication Services,” Automatika, vol. 51, no. 3, pp. 275–283, 2010.

[14] A. Backiel, B. Baesens and G. Claeskens, “Predicting time-to-churn of prepaid mobile telephone customers using social network analysis”, Journal of the Operational Research Society, pp. 1135-1145, 2017.

Marin Mandic was born on 1977. in Mostar, Bosnia and Herzegovina. He recived his Graduate degree of Information System in 2005 at Faculty of Organization and Informatics in Varazdin. Since 2008. he has been employed in company HT Eronet Mostar. Since 2016. he has been a Ph.D student at Faculty of Mechanical Engineering, Computing and Electrical Engineering, University of Mostar. His research area is Data mining, Data Warehouse and Social Network Analysis.

Goran Kraljevic was born on 1977. In Mostar, Bosnia and Herzegovina. He received his B.Sc. degree in Computing in 2000. from the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia. He received his PhD degree in Computer Science in 2012. from the Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, University of Split, Croatia. Since 2003. he has been employed in company HT Eronet Mostar. Since 2004. he has worked as an external Assistant Professor at the Faculty of Mechanical Engineering, Computing and Electrical Engineering, University of Mostar. His research areas of interest include Data mining and knowledge discovery, Data Warehousing and Business Intelligence systems.

Ivan Boban was born on 1987. in Mostar, Bosnia and Herzegovina. He received his Graduate Engineer Degree in 2014. Since 2016. he has been a Ph.D student at Faculty of Mechanical Engineering, Computing and Electrical Engineering, University of Mostar. Since 2012. he is also a Business and Analytical Information Systems Associate at JP Croatian Telecommunication d.o.o. Mostar. His research interests include Information Retrieval and Sentence Retrieval.