EAI Endorsed Transactions
on Energy Web
Research Article EAI.EU

Improvised_XgBoost Machine learning Algorithm for Customer Churn Prediction

Swetha P[1]*, Dayananda R B[2]

[1]Research Scholar, Visvesvaraya Technological University, Assistant Professor, Department of CSE, KS School of Engineering and Management, Bangalore, India.
[2]Professor, Department of CSE, KS Institute of Technology, Bangalore, India.

Abstract

The Customer retention has become one of the major issues for the service-based company such as telecom industry; where predictive model to observe customer, behavior is one of the efficient methods in the customer retention process. In this research work, Improvised_XGBoost churn prediction model with feature functions is proposed, the main aim of this model is to predict the customer churn rate. Improvised_XGBoost algorithm is a feature-based machine learning classifier which can be used for the complex dataset. At first, feature function is introduced then loss function is formulated and minimized through iterative approach, later combined with XG_Boost approach it possesses better efficiency. The main feature of Improvised_XGBoost algorithm is that it handles the unstructured dataset attributes efficiently, further feature function combined with XG_Boost. Furthermore, the proposed model is evaluated through various performance metrics such as accuracy, precision and recall. Our model also throws light on identifying the correctly and incorrectly classified instances on South Asia GSM (Global System for Mobile Communication) service provider. The results through the comparative analysis, our model outperforms the other state-of-art technique.

Keywords: customer churn prediction, improvised-XG boost, telecommunication, prediction model.

Received on 07 May 2020, accepted on 19 May 2020, published on 02 June 2020

doi: 10.4108/eai.13-7-2018.164854

*Corresponding Author. Mail: Shwetha6600@gmail.com

1. Introduction

Telecom industry has observed enormous growth over the last decade, it is a leading business among the subscription-based in last two decades. Further, this will prove through the fact that the number of mobile users has increased over more than 8 billion all over the world [1] [2], hence this causes an avalanche of data in a daily basis. These data are divided into two categories i.e. real-time and batch mode. Further data includes VOIP data, hardware, customer profile, network performance, network monitoring, graphical user data, and mobile network usage, and user click stream, call detail records. etc. In this Big data era, these data can be categorized into velocity, variety and volume [3] [4]. Moreover, some of the developing nations have more telecom subscribers than its inhabitant. However, in the last few year's countries like India have seen a dramatic change in the telecom industry and several telecom services suffered from a huge loss which causes them to shut down their services. Moreover, this took place due to the different strategies adopted by a different company, hence considering such scenarios retaining their customers has become the top priority of telecom companies for maintaining a sustainable profit. Further, it has been observed that retaining a customer is much more important and beneficial than getting new customer [5]; as getting a new customer costs nearly 5 to 6 times more than retaining a customer and also periods is required to develop the loyalty of the particular customer with the telecom companies. Customer loyalty is built based on the satisfaction factor of the customer with the telecom services and how efficiently the companies have matched

their demands. Hence, this requires more efforts to gain a new customer; however, retaining a customer does not involve such a long process and also it does not require other expenses such as additional marketing. Loyal and long-term customers will generate high profits for the companies as they are not vulnerable and do not get easily attracted by other telecom services in comparison with the new customers. In general, customer churn takes place either voluntarily or involuntarily, when the particular customer is not satisfied with the services offered by telecom service providers. Such unsatisfied customers start switching to different telecom service providers network. There can be multiple reasons for churning such as customer service, cost of data, network issue and many others. Thus, churn prediction is a top priority for any company to stay alive in the market and which can be easily achieved and monitored through CRM (Customer Relationship Management) [6]-[7]. The main aim of the CRM team is to maintain and monitor the likes and dislikes of each customer along with his call log details. Based on the valuable information available with the CRM team, the service providers can design and develop the retention policies individually for each customer. These policies should have a high accuracy rate and model should be capable of identifying the customers who are willing to churn now and later with the reason behind their churning. Thus in a few last types of research, more focus was towards developing the prediction model. In general, the prediction model involves a huge amount of customer data, so data mining and machine learning domains were used for analyzing such large data. These techniques are involved in identifying the customer churn rate, further classification approach is used for utilizing the customer characteristics. Classification approach will achieve the customer retention policies through the pattern extraction of data mining technique and is combined with data and decision-makers through the CRM analyzer.

In past several prediction models have been developed such as neural networks and decision trees were used for developing the absolute prediction model [8]; however, complexity leads to varying and was less accurate. Hence, different variants of Rot-boost and rotation forest is used for high accuracy [9]. [10] Developed hybrid approach, which combines the k-means and rule technique for attaining high accuracy, similarly [11] presented GA-based neural network approach for increasing the accuracy, these approaches were either on individual-based or ensemble classification. Further, these classification methods use the feature extraction and sampling method. [12] Analyzed the impact of sampling in churn prediction and also under sampling were explored widely here it is observed that few researchers used the CUBE sampling technique, however, these sophisticated sampling did not produce the desired result. Other approaches such as [13] used a Bayesian approach for feature extraction. Moreover [14]developed a churn prediction model using the data uncertainty and [15] used PSO approach; these were carried out through comparative analysis of ANN and decision trees and analysis indicates that decision trees perform far better than the neural network, these were extensively discussed in [16], further [17] extended this work and aimed at finding the customer loyalty. Moreover, the two-step approach is used, the first step is related to RFM which divides the feature into four clusters and the second step includes churning the data.

Few researchers have focused on the particular features to induce classification algorithm for churn prediction, these features included line information, service records, account information, call details and payment records; along with this decision trees, multi-perceptron and support vector machine was used. Considering these features [18]-[19] developed multi-objective feature selection which was NSGAII based, other paper-like [20] relied on Bayesian Belief Network for extracting these features.

[21] Developed another intelligent model of data mining for churn prediction with good accuracy, here author applied PCA (Principal Component Analysis) for data dimension reduction. For classification three algorithm Bayes network, SVM and Neural Network were used; further 3333 customers sample were used along with 21 features churn parameter were used as yes or no.

[22] Was based on Neural Network for solving the churn prediction, here large dataset of Chinese telecom industry was used which consist of 5.23 million customers and achieve accuracy of 91.1%. [23] Proposed integration of Ada-boost and genetic programming for churn prediction, they used the two standard dataset cell 2 cells and achieved an accuracy of 91.1% for orange and 63%for cell2cell dataset. [24] Provided a comparative analysis of six different sampling technique developed for churn prediction and comparison analysis observed that genetic algorithm was better than other algorithms. Considering the above analysis of existing model, it is observed that machine learning algorithm are best to consider for further research. In the above existing method, some method is either complex or it fails with another dataset; the main drawback of these models is that they are incapable of handling unbalance and large dataset. Hence considering these drawbacks we would be proceeding our research in a designing model that can efficiently handle dataset with better accuracy; the proposed model is discussed in the next section.

This research work is designed like any standard research work, here the first section starts with background stats of telecom industry and further we discuss concern rising in the telecom industry; later we discuss a various existing model for churn prediction and highlight their shortcomings. Further, in the same section, we discuss the motivation and contribution of research work. The second section discuss proposed Improvised_XGBoost and present their mathematical model; the third section evaluates the proposed model considering various performance metrics.

1.1 Motivation and contribution of the research work

Churning could happen due to various reasons; In general, the service provider like post-paid customers does not bound prepaid customers. Hence, pre-paid customers are more willing to churn than post-paid customers. These customers can churn at any time irrespective of their time bonding with service providers. Further, these churning factors impact directly on the company's reputation which in turn cause an impact on brand image. However, a loyal customer gets rarely affected by the other telecom service provider and he or she may refer to friends, colleague and their family. Moreover, the telecom service provider needs to consider the shift in policy if the customers drop to the certain level, this might cause a huge financial loss for telecom companies, and hence it is very important for retaining the customers. Retaining customers is a tough task considering the various scenarios, hence in this research work we have developed a methodology for churn prediction based on the XG_Boost algorithm, XG_boost is machine learning classifier. The major contributions of this research work are given as follows:

- This research works proposed an Improved XG_Boost methodology for churn prediction considering the large dataset.
- Improvised_XGBoost algorithm is a feature-based classifier which classifies unstructured data and predicts the churn rate.
- Improvide_XGboost is the combination of feature function with the XG_Boost to understand the tree problem and for efficient classification of a complex dataset.
- The further loss function is minimized iteratively to achieve higher accuracy, besides another advantage of Improvised_XGBoost is that it deals with various attributes very smoothly and efficiently.
- Improvised_XGBoost is evaluated by considering the important measuring metrics such as precision, recall and accuracy on the standard dataset available.

2. Proposed Methodologies

In this section, we have designed and developed an Improvised_XGBoost prediction model for customer churn prediction along with feature function development for handling large dataset with complex attributes, which in turn predicts the churn rate.

2.1 Improvised_XGBoost prediction model

XG-Boost is nothing but EGB (Extreme Gradient Boosting); it is designed for better performance and high speed. In general, XG-Boost is a dominated ensemble supervised learning technique, which will increase the computing power limits of boosted decision trees. Moreover, in here we propose improvise_XGBoost which adds the feature function; feature function added with XGBoost gives the advantage of handling the imbalance and complex data. Improved-XGBoost classifier comprises various gradient boosted trees, which are designed to achieve high results in prediction. These trees are built one after the other sequentially which reduces the error defined in the previous tree and builds a corrected new tree each time. As XG-Boost classifier provides the better results in terms of prediction; however, it requires more time to train the data in an iteration process, however, Improvised-XGBoost provides better training model with greater accuracy and also focus on the problems of tree learning. Moreover, to achieve better accuracy feature function is introduced, feature function along with XG_Boost provides a better understanding of tree learning process and achieve higher performance metrics.

Figure 1 shows the process flow of Improvised_XGBoost classifier; just like another machine learning classifier, Improvised_XGBoost comprises four-step process flow as shown in four blocks of Figure1. In the very first block, we access the data and which is taken as input to the given model. In the second blocks of Data wrangling, where the data is converted and mapped into the required format for analysis purpose. The third block involves the training of data through our Improvised_XGBoost classifier and results are tested for better accuracy. Followed by the fourth block, which involves the prediction, where acquiring insights are carried out and the outcome is predicted and evaluated on several metrics.

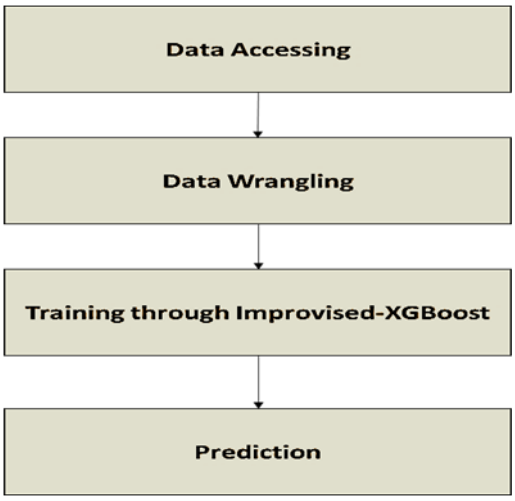

Figure 1. Improvised_XGBoost classifier process flow

2.2 Feature-based classification model

In this section, we focus on developing feature function by considering the feature as a primary component and later loss function is formulated and optimized through an

iterative approach. For the given training dataset along with given samples, the prediction is depicted through the below equation.

$$\overline{\mathbb{W}}_j = \mathbb{D}(\mathbb{V}_j) = \sum_{m=1}^{m} d_m(\mathbb{V}_j) \qquad (1)$$

$\mathbb{V}_j$ Denotes the jth sample given in training set, further, it is observed that$d_m \in \zeta$; ζindicates the decision trees set which can be represented through the below equation.

$$\zeta=\{d(\mathbb{v}) = \mathbb{u}_{q(\mathbb{v})}\} \qquad (2)$$

In the above equation, $\mathbb{u}$ indicates leaf features and q represents the structure parameter, each decision trees $d(\mathbb{v})$ corresponds to the features of the leaf. $\overline{\mathbb{W}}_j$ Indicates prediction. The purpose of this function is to reduce the loss function.

$$\mathbb{N}(\mathbb{D})=\sum_m \Xi\,(d_m) + \sum_j n\left(\overline{\mathbb{W}}_j, \mathbb{W}_j\right) \qquad (3)$$

Equation 3 represents the ensemble tree model. The above equation comprises two terms; the first term i.e. represents the penalty function and the second term indicates the loss function; $\left(\overline{\mathbb{W}}_j, \mathbb{W}_j\right)$ indicates the loss function between the predicted label and real label data. Two parameters i.e. Ξ and Θ are used for controlling the tree complexity. In improvised_XGBoost. Penalty term can be formulated through the below equation.

$$\Xi(e_m) = \Theta\|\mathbb{u}\|^2 + \Gamma\mathbb{R} \qquad (4)$$

Cross entropy loss is used to train the sample $\mathbb{v}$ with id $\mathbb{m}$, hence the loss function is formulated as below.

$$\mathbb{m}\left(\overline{\mathbb{W}}_j, \mathbb{W}_j\right) = -\sum_j \mathbb{w}(i) log\, \overline{\mathbb{W}}_j = -log\overline{\mathbb{W}}_j(\mathbb{m}) \qquad (5)$$

$\mathbb{W}(i)$is ith class of $\mathbb{W}$, i represents the probability of getting predicted, $\overline{\mathbb{W}}(\mathbb{m})$ indicates $\mathbb{m}th$ class of output $\mathbb{W}$. The further loss function is minimized through optimizing the function iteratively, the iteration takes place till rth value.

$$\mathbb{N}^r = \sum_{j=1}^{l} \mathbb{m}\left(\mathbb{w}_j^{(r-1)}, \mathbb{w}_j\right) + d_r\left(\mathbb{v}_j\right), \mathbb{w}_j + \Xi(d_m)\Big) \qquad (6)$$

Once the problem is defined it is very difficult to enumerate the possible structure of r, hence an algorithm is designed which starts from the single leaf and keeps on adding the branches and $d_r$is calculated. Equation 6 can be further minimized through the below algorithm as shown in following table. Table 1 presents the steps for finding the optimized split.

Table 1. Improvised_XGBoost algorithm

Step1:	Set of instance and set of features are given as input
Step2:	Consider gain as null
Step3:	$\sum_{j\in\mathbb{J}} \mathbb{F}^j \Longrightarrow \mathbb{F}$
Step4:	$\sum_{j\in\mathbb{J}} \mathbb{G}_j \Longrightarrow \mathbb{G}$
Step5:	For j= 1 to n do
Step6:	$\mathbb{F}_M = 0$ and $\mathbb{G}_M = 0$
Step7:	For k in sorted $(\mathbb{J},\ by\ \mathbb{W}_{ij})$ do
Step8:	$\mathbb{F}_M + f_j \Longrightarrow \mathbb{F}_M$
Step9:	$\mathbb{G}_M + g_k \Longrightarrow \mathbb{G}_M$
Step10:	$\mathbb{F} - \mathbb{F}_M \Longrightarrow \mathbb{F}_q$
Step11:	$\mathbb{G} - \mathbb{G}_M \Longrightarrow \mathbb{G}_q$
Step12:	$maximum(\frac{\mathbb{F}_M^2}{\mathbb{G}_M+\Xi} + \frac{\mathbb{F}_M^2}{\mathbb{G}_q+\Xi} - \frac{\mathbb{F}^2}{H+\Xi}, SC) \Longrightarrow SC$
Step13:	End of for loop(step7)
Step14:	End of for loop(step8)
Step15:	Output max score

Further, this algorithm (table1) explores all the splits through the continuous feature extraction in an efficient manner which is achieved by analyzing the feature values and sorting the data according to it. Once after we get the output as the prediction score then we have evaluated the proposed model to prove the efficiency of our prediction model.

3. Performance Evaluation

In this section, we have evaluated our proposed model Improvised_XGBoost and which is carried out through the system configuration of 8 GB RAM, with 1TB hard-disk and 2 GB NVidia graphics card; further windows 10 platform is considered. Moreover, R programming language is used for programming and visual basic 2017 is used as the programming platform.

3.1 Dataset description

To evaluate the Improvised_XGBoost algorithm we have considered standard dataset, which is obtained from South Asia GSM (Global System for Mobile), further, this dataset comprises 29 distinctive features along with 64,107 instances given in table 2.

These features are numerical data; moreover, these data have been extracted through CDR (Call Detail Record). Further, the dataset is labelled in two distinctive class namely "T" and "F", T indicates true customers and it is 30% of labelled data, F indicates false customers and it is of 70% of labelled data. Moreover, true customers are the churners and false customers are non-churners; further 29 attributes are classified into three types which include financial information attributes, marketing-related attributes and call behaviour. Moreover, these attribute selections depend on feature selection technique which identifies the effective dataset [25].

Table 2. Dataset Description

Dataset	Instance	Attributes	Target Class
South Asia GSM telecom	64,107	29	Two class classification, where F is non-churn customer and T is churn customer

3.2 Performance Metrics

In this section, we discuss the different performance metrics used for evaluation of Improvised_XGBoost, which proves the efficiency of the model. This research work considers three important metrics namely accuracy, precision and recall. Moreover, existing model [26] has considered various algorithms for comparison.

3.2.1 Accuracy

Accuracy is defined as the closeness of generated value to the defined value, in here it is correctly classified value. Accuracy identifies the two types of classification i.e. correctly classified instances and incorrectly classified instances. In general, the correctly classified instance is computed through the ratio of the addition of true positive and true negative with the addition of true positive, true negative, false positive and false negative. Accuracy is the major metric for evaluating any method, Table 2 shows the comparison of various methodologies with correctly and incorrectly classified values and the comparative analysis of those methodologies indicates that Improvised_XGBoost model achieves correctly classified accuracy value of massive 99.41 and incorrectly classified value of 0.59 %. The graphical representation of comparison is shown in figure2.

Table 3. Comparison of various methodology on the correct and incorrect classification

Methodology	correctly classified	incorrectly classified
Random Forest	88.63	11.37
Bagging + Random	88.61	11.39
J48	88.58	11.42
ASC (Attribute Selected Classifier)	88.34	11.66
Random Tree	84.34	15.66
AdaBoostM1	83.95	16.05
Multilayer Perceptron	82.04	17.96
LWL	81.59	18.41
IBK	80.37	19.63
Classifier +Decision Stump	70.98	29.02
Logistic Regression	70.98	29.02
Naïve Bayes	47.63	52.37
Improvised_XGBOOST	99.41	0.59

3.2.2 Graphical Representation

Fig 2. Shows the comparison of a various prediction model with improvised_XGBoost, in here x-axis depicts methodology y-axis indicates correctly classified and incorrectly classified. Moreover, in the graph, we observe that our methodology achieves higher correctly classified and lower incorrectly classified.

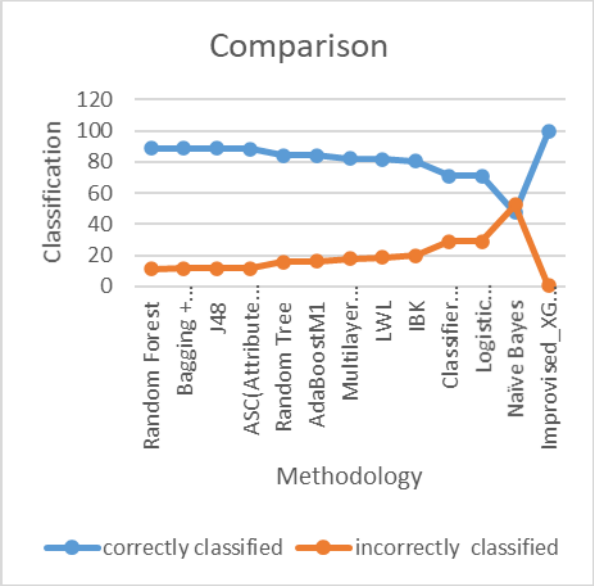

Figure 2. Graphical comparison of various methodology.

3.2.3 Precision and Recall

In this section, Improvised_XGBoost is evaluated by considering another two important metrics precision and recall;

Precision is one of the important performance metrics and it is defined as the percentage of whole relevant outcomes correctly classified by the classifier.

Furthermore, comparative analysis is carried out with various state-of-art technique and the same is given in Table 3. Precision is also known as positive predictive value and it shows that how much part of data predicted is positive. In general, precision is the ratio of true positive to the addition of true positive and false positive. From the table3, we observe that Improvised_XGBoost model achieves 0.9944, which is marginally higher than any other model. Similarly, Recall is defined as the probability where the relevant instances are chosen, low recall value indicates the number of false negatives. In general, recall is defined as the ratio of true positive and the sum of a truly positive and false negative. From table 3 we observe that Improvised_XGBoost achieves recall value of 0.9994, which is marginally high when compared to any other model.

Table 4. Precision and Recall comparison of various methodologies

Methodology	Precision	Recall
Random Forest	0.893	0.888
Bagging + Random	0.883	0.881
J48	0.893	0.887
ASC (Attribute Selected Classifier)	0.902	0.888
Random Tree	0.844	0.843
AdaBoostM1	0.839	0.835
Multilayer Perceptron	0.821	0.822
LWL	0.806	0.812
IBK	0.805	0.81
Logistic Regression	0.49	0.7
Naïve Bayes	0.715	0.473
Improvised_XGBoost	0.9944	0.9994

4. Conclusion

Customer churn management has become one of the mandatory elements of all telecom companies. In this era, huge competition exists between different service providers however existing churn prediction model may not work well due to inefficient data management and analysis techniques. Our research work focused on developing a new churn prediction model named Improvised_XGBoost. This model is a feature aware model, which handles large dataset and its attribute feature function supports it to handle various attributes efficiently. This algorithm is designed and developed to achieve better accuracy. Further, our proposed model is evaluated by considering the large dataset with fair instances. The proposed model achieves massive accuracy of 99.41% and precision value of 99.44 % and recall value of 99.94%. Although our model achieves the nearer to absolute accuracy, still few areas need to be focused and needs to be evaluated by considering different dataset which would be carried out in our future work. Churn prediction is an extensive research area; hence this work can be considered for further research in designing prediction model.

References

[1] P. T. Kotler, Marketing Management: Analysis, Planning, Implementation and Control. London, U.K.: Prentice-Hall, 1994.

[2] S. Babu, D. N. Ananthanarayanan, and V. Ramesh, ``A survey on factors impacting churn in telecommunication using datamininig techniques," Int.J. Eng. Res. Technol., vol. 3, no. 3, pp. 1745_1748, Mar. 2014.

[3] P. Zikopoulos, C. Eaton et al., Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. McGraw-Hill Osborne Media, 2011.

[4] D. Sipus, "Big data analytics for communication service providers," in 39th IEEE Int. Conv. Information and

Communication Technology, Electronics and Microelectronics, May 2016.
[5] L. Goleniewski and K. W. Jarrett, Telecommunications Essentials: The Complete Global Source, 2nd ed., K. W. Jarrett, Ed. USA: Addison Wesley Professional;, 2006.
[6] M. Kaur, K. Singh, and N. Sharma, ``Data mining as a tool to predict the churn behaviour among Indian bank customers," Int. J. Recent Innov.Trends Comput. Commun., vol. 1, no. 9, pp. 720_725, Sep. 2013.
[7] V. L. Miguéis, D. van den Poel, A. S. Camanho, and J. F. e Cunha, ``Modeling partial customer churn: On the value of _rst product-category purchase sequences," Expert Syst. Appl., vol. 12, no. 12, pp. 11250_11256, Sep. 2012.
[8] Shin, Y.H., David, C.Y., Hsiu, Y.W.: Applying data mining to telecom churn management. Expert Syst. Appl. 37, 3665–3675 (2006).
[9] Bock, K.W.D., Van den Poel, D.: An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction.Expert Syst. Appl. 38, 12293–12301 (2011).
[10] Huang, Y., Kechadi, T.: An effective hybrid learning system for telecommunication churn prediction. Expert Syst. Appl. 40, 5635–5647 (2013).
[11] Pendharkar, P.C. Genetic algorithm-based neural network approaches for predicting churn in cellular wireless network services. Expert Syst. Appl. 36, 6714–6720 (2009).
[12] Burez, J., Van den Poel, D.: Handling class imbalance in customer churn prediction. Expert Syst. Appl. 36, 4626–4636 (2009).
[13] Kisioglu, P., Topcu, Y.I.: Applying Bayesian belief network approach to customer churn analysis: a case study on the telecom industry of Turkey. Expert Syst. Appl. 38, 7151–7157 (2011).
[14] A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo, and S. Anwar, ``Customer churn prediction in telecommunication industry using data certainty," J. Bus. Res., vol. 94, pp. 290_301, Jan. 2019.
[15] J. Vijaya and E. Sivasankar, ``Anefcient system for customer churn prediction through particle swarm optimization-based feature selection model with simulated annealing," Cluster Comput., pp. 112, Sep. 2017.
[16] 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, Mar. 2012.
[17] A. T. Jahromi, M. Moeini, I. Akbari, and A. Akbarzadeh, ``A dual-step multi-algorithm approach for churn prediction in pre-paid telecommunications service providers," J. Innov. Sustainab., vol. 1, no. 2, pp. 2179_3565, 2010.
[18] Huang, B.Q., Kechadi, T.M., Buckley, B., Kiernan, G., Keogh,n E., Rashid, T.: A new feature set with new window techniques for customer churn prediction in land-line telecommunications. Expert Syst. Appl. 37, 3657–3665 (2010)
[19] Huang, B., Buckley, B., Kechadi, T.M.: Multi-objective feature selection by using NSGA-II for customer churn prediction in telecommunications. Expert Syst. Appl. 37, 3638–3646 (2010)
[20] Kisioglu, P., Topcu, Y.I.: Applying Bayesian belief network approach to customer churn analysis: a case study on the telecom industry of Turkey. Expert Syst. Appl. 38, 7151–7157 (2011)
[21] Brandusoiu I, Toderean G, Ha B. Methods for churn prediction in the prepaid mobile telecommunications industry. In: International conference on communications. 2016. p. 97–100.
[22] He Y, He Z, Zhang D. A study on prediction of customer churn in fixed communication network based on data mining. In: Sixth international conference on fuzzy systems and knowledge discovery, vol. 1. 2009. p. 92–4.
[23] Huang F, Zhu M, Yuan K, Deng EO. Telco churn prediction with big data. In: ACM SIGMOD international conference on management of data. 2015. p .607–18.
[24] Amin A, Anwar S, Adnan A, Nawaz M, Howard N, Qadir J, Hawalah A, Hussain A. Comparing oversampling techniques to handle the class imbalance problem: a customer churn prediction case study. IEEE Access. 2016;4:7940–57.
[25] https://www.kaggle.com/mahreen/sato2015.
[26] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam and S. W. Kim, "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector," in IEEE Access, vol. 7, pp. 60134-60149, 2019.