Churn Prediction Models Using Gradient Boosted Tree and Random Forest Classifiers

1st Yu Yu Win
University of Computer Studies, Mandalay
Mandalay, Myanmar
yuyuwin2@ucsm.edu.mm

2nd Cing Gel Vung
University of Computer Studies, Mandalay
Mandalay, Myanmar
cinggelvung@ucsm.edu.mm

***Abstract*—In the era of a competitive market, every organization has been used a lot of marketing techniques to maximize their profit and to preserve the existing flow of customer relationships with the firm. The cost of attracting a new customer incurs more times than retaining existing ones. Thus, customer relationship management (CRM) analyzers try to know the behavior of customers and find the causes of a customer churning. To produce a list of telecom customers who likely to churn in the future, this paper presents the two churn prediction models using wrapper-based Forward Feature Selection (FFS) with Gradient Boosted Tree and Random Forest classifiers. This work analyzes the FFS with five comparative classifier models based on the telecom data using the KNIME analytics platform and deploys the two most accurate models with a new dataset to predict the future churn. Our models achieve the accuracies of 96.2% and 96.89% respectively.**

Keywords— CRM, KNIME, customer churn, prediction, telecom, machine learning

I. Introduction

The telecom companies undergo several competitive environments in marketing strategies. The telecom providers always take care of their services to satisfy their customers, possibly cheaper prices, and better services. Customer churn happens when a vast percentage of clients are not satisfied with the services of any telecom company. It results in the service migration of customers who start switching to other service providers [1]. Bart Baesens [2] divided the customer churn into four categories.

1) Active Churn – Customers stop the relationship with the firm.
2) Passive Churn – Customers decrease the intensity of the relationship with the firm.
   For example: by decreasing products or services used.
3) Forced Churn – The Company stops the relationship with the customer because he or she has been engaged in fraudulent activities.
4) Expected Churn – Occurs when the customers no longer need the products or services, for example, baby products.

In these types of churn, this paper is to predict active churn. Customer management is the most effective factor to steer the success of the company. Customer churn is a helpful metric for detecting the weakness of a company [3]. One way to ensure effective customer relationship management is through carefully monitoring churn. With the rapid growth of telecommunication companies, customers can easily switch to other telecom operators. One fundamental factor of telecom companies is to predict customers who are likely to churn. As a result, churn prediction is a vital business metric as well as one of the vastly important machine learning applications in telecom companies.

Data mining aims to discover hidden patterns or trends in the large dataset [4]. There are many techniques in data mining and the classification model is extensively used in many areas to find trend analysis and future planning. A classification model is also known as supervised learning which analyzes the training data and produces a correct outcome from labeled data. It is a popular model in data mining for predicting among other models.

In this study, a telecom dataset is used to classify churn or not churn. For training, five classification models naming Decision Tree (DT), Logistic Regression (LR), Gradient Boosted Tree (GBT), Random Forest (RF), Naïve Bayes (NB), have been used. This paper builds the models and uses an unseen dataset to predict future churn. Moreover, the system presents the reason why the customers are likely to churn, so these will be useful to choose targeted customers in retention strategies.

The rest of this paper is organized as follows. Section II presents the related work of machine learning for churn prediction. Section III describes the data preprocessing stage to remove the dummy data and feature selection to choose the important ones. The churn prediction model is presented in Section IV. The experimental result for churn prediction is presented in Section V and the conclusion is in section VI.

II. Related Work

As the number of modern telecommunication providers is increasing, the number of customers who likely to churn is very increasing. Hence, gaining market share is harder. In business, future prediction using machine learning is popular. The existing works of literature for churn prediction are vast amounts.

In paper [5], the "KDD cup 2009" dataset is used to predict customer churn for Customer Relationship Management. The authors examined the telecommunication dataset using Convolutional Neural Network (CNN) which is a class of Deep Learning approach and Feedforward Artificial Neural Network. CNN applied a convolutional layer that filters input values producing feature maps for that filter. The authors tried to tune the network layers. Their final accuracy is more than 95%.

In paper [6], churn prediction is analyzed in Mobile Games. This paper was a little different from another one. It used not only a churn vector to get the prediction accuracy but also strategic promotions to minimize game user attrition. It examined the applicability of different algorithms like Lasso, Support Vector Machines, Recurrent Neural Network (RNN), Decision Tree, Random Forest, and Gradient Boosted

Machine. The authors found not only R2 values of Regression but also Accuracy of Classification. Among all models, RNN is most accurate and the accuracy of classification models increased from 1% to 3%.

Anurag Bhatnagar and Sumit Srivastava [7] analyzed the churn dataset on the performance of the Hoeffding and Logistic Algorithm. They have designed the models for classifying churn under unsupervised machine learning. They used one of the most powerful tools Weka. To classify the data, a classifier performance evaluator was used. The final inference of their system was that the logistic algorithm outperformed the Hoeffding Tree algorithm for customer churn prediction.

With the Just-in-time churn prediction analysis [8], the authors examined the telecom dataset using the cross-company concept (i.e., when one company (source) data as a training set and another company (target) data was considered for testing purpose) with three experiments. Their proposed method was intended for a situation where the company is newly established or have lost the historical data relating to the customers. Their first experiment was done without data transformation methods (original data) and AUC was obtained 0.499. The second experiment was performed with the data transformation method by considering the log method on both datasets and AUC is 0.491. The final experiment was similar to the second one and for data transformation, the rank method was used. The final AUC was 0.593. The empirical results showed that the rank method outperformed as compared to the log method. All three experiments were processed using Naïve Bayes Algorithm.

Using supervised machine learning techniques, two classification models, KNN and Logistic Regression were implemented using the python platform in [9]. The authors compared the performance of these two models and presented the pros and cons of using the KNN models. The confusion matrix is observed that KNN is a better approach to predict customer churn over Logistic Regression. KNN is more accurate 2.0% than Logistic Regression with an accuracy of 88.5% and 86.5%. Many researchers used various methods of churn prediction and many machine learning platforms such as Weka, KNIME, and used python with Jupyter Notebook, and spyder on Anaconda.

III. Data Preprocessing

In data preprocessing, there are two steps, data cleaning, and feature extraction. At first, we introduce the KNIME (Konstanz Information Miner) analytics platform is a free and open-source data analytics, reporting, and integration platform. Its purpose is for fast, easy, and intuitive access to advanced data science, helping organizations drive innovation [10]. In KNIME, the user selects the basic element called a node. Each node in KNIME has its functionality, based on the needs of the user, can configure them as per the application [11]. The data in the study is available from the Kaggle website[1]. It includes 598 "churn" records and 3652 records are "no churn". The dataset contains 20 features including class label, churn. Table I presents the description of the attributes.

[1] https://www.kaggle.com/c/customer-churn-prediction-2020/data

TABLE I. Description Of Attributes

No	Name	Data Type
1	Id	string
2	State	string
3	Accountlength	integer
4	Area code	string
5	International plan	string (yes, no)
6	Voice mail plan	string (yes, no)
7	Number vmail message	integer
8	Total day minutes	double
9	Total day calls	integer
10	Total day charge	double
11	Total eve minutes	double
12	Total eve calls	integer
13	Total eve charge	double
14	Total night minutes	double
15	Total night calls	integer
16	Total night charge	double
17	Total intl minutes	double
18	Total intl calls	integer
19	Total intl charge	double
20	Churn (class label)	string (yes, no)

This paper used data cleaning with manual selection of nodes in KNIME to filter the missing and dummy values. The column "id" is removed manually. Data produced by these nodes are examined with the Forward Feature Selection (FFS) method which is one type of wrapper-based methods. It is an iterative approach. Naïve Bayes classifier is used to test the model improvement. We start with having no feature in the model. In each iteration, the feature that improves the model the most is added to the feature set [10]. In similarly, we keep adding the feature which best improves our model till an addition of a new variable does not improve the performance of the model. In FFS, the Naïve Bayes algorithm uses all the features in the input dataset and selects the subset of features that are best for model construction. The workflow of this method includes learner node and predictor node between Feature Selection Loop Start node and Feature Selection Loop End node. It is shown in Fig1.

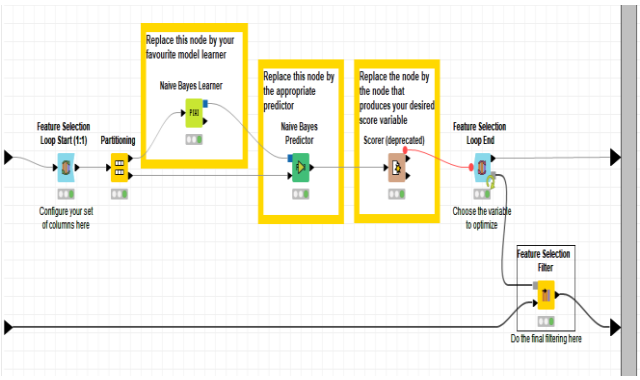

Fig 1. Forward Feature Selection Workflow

In Fig 1, the dataset is split for training and testing. After iteration, the result is presented with the Feature Selection Filter node. There are 17 rules with an accuracy of 83.2% to 90.8% using 15 and 16 features respectively. Each accuracy has each of the numbers of features. To find a higher performance model, we compare accuracy and Cohen's Kappa using 17 features (with accuracy 90.8% of feature selection) and 16 features (with accuracy 87.3%), it is shown in Table II. We build the models using the features with the highest accuracy of 90.8% from feature selection. These 17 features are all features except id, state, and Accountlength

including the class label. The accuracies of the models using FFS increase almost 2% than those without using feature selection.

TABLE II. DESCRIPTION THE ACCURACY OVER FORWARD FEATURE SELECTION USING 17 AND 16 SELECTED FEATURES

Machine Learning Algorithm	All Features		17 Features (A=90.8%)		16 Features (A=87.3%)	
	A	CK	A	CK	A	CK
Decision Tree	92.6	65.2	94.0	80.3	94.4	73.7
Logistic Regression	86.5	23.8	87.3	24.5	86.9	25.3
Gradient Boosted Tree	94.5	83.9	96.2	81.8	95.3	78.5
Random Forest	95.2	77	96.89	85.1	94.8	74.3
Naïve Bayes	89.3	48.1	90.08	43.7	87.3	32.3

A=Accuracy, **CK**=Cohen's Kappa

IV. CHURN PREDICTION MODELS

Five machine learning models used to train the dataset are Naïve Bayes (NB), Random Forest (RF), Decision Tree (DT), Gradient Boosted Tree (GBT), and Logistic Regression (LR). The training workflow is shown in Fig 2. The dataset received from the FFS node is 4250 records with 17 columns. The partitioning node splits these data with an 8:2 ratio for training and testing data. Training data is trained on DT learner, LR learner, GBT learner, RF learner, and NB learner node each. Each learner node produced its related models. Each model is connected with each related predictor nodes. A predictor node has two inputs: model and testing dataset, and one output: classified dataset with percentages of churn or non-churn and prediction are yes or no. Testing data of 20% is used in each predictor node.

Using the column filter node, the probability (churn=yes), prediction (churn) columns are extracted without original data columns. By using column rename node, each prediction is renamed with each model name to use in Binary Classification Inspector Node. All columns are combined with the "column appender" node and also with testing data.

The binary classification inspector node produces a complex view made of four different charts to compare, optimize, and select predictions of different binary classifiers. It compares several binary classifier machine learning models predicting the same target on the same test data using performance metrics and ROC curves. It optimizes a model by finding the best threshold given a performance metric of user choice. Interactively select a given type of prediction (e.g. True Positives) of one of the models and export them at the output of the node.

The user journey model AUC to select the best model via the Model's Statistics bar chart and the Model's ROC Curves chart. We can change the threshold from its initial value either manually, via the threshold slider. Inspect the confusion matrix to assess the gravity of the misclassification, give the associated probability confidence of the model on the Classification Distribution chart. Thus, the binary classification inspector node is very easy for comparison of the prediction models.

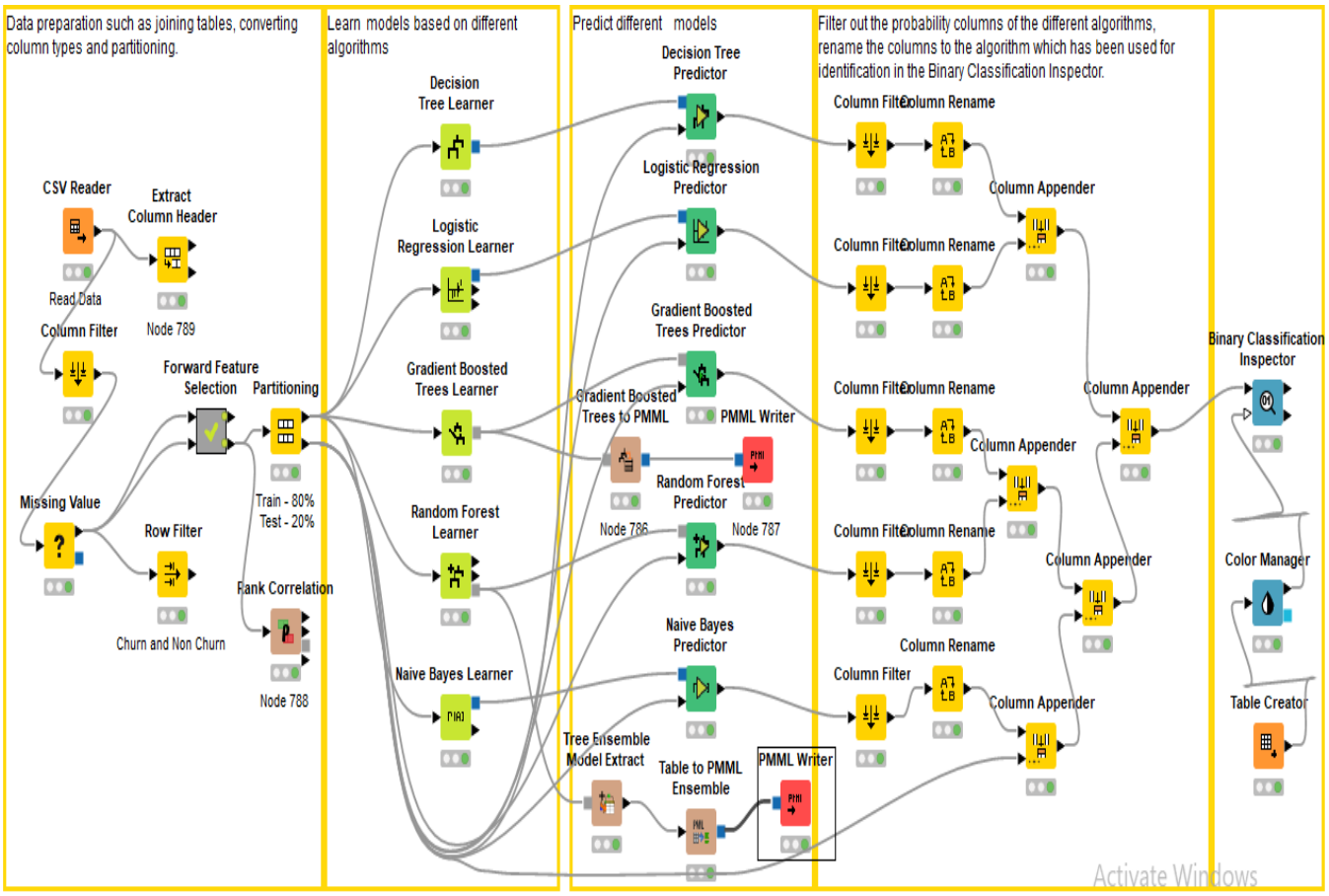

Fig 2. Churn Prediction Model Workflow in KNIME

V. Experimental Results

The objective under the problem is to select the model that yields good classification accuracy with maximum precision and Cohen's kappa value. To increase the accuracy of each model, we can adjust the threshold value. This system chooses the threshold value of 0.5, the accuracies of all models produced by the binary classification inspector node is given in Table II and the visualization is shown in Fig 3.

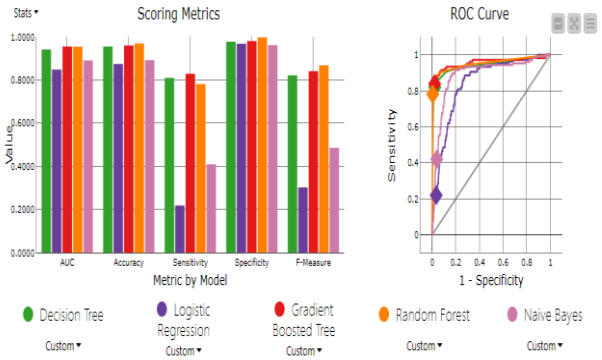

Fig 3. Visualization of Accuracy of all Models

According to the accuracy table and visualization, GBT and RF outperform other models in the mean of not only accuracy, F-Measure but also Cohen's Kappa value, AUC. The false-positive and false-negative numbers of GBT are 18, 15 and those of RF are 23 and 2. These models achieve less miss classification (numbers of false positive and false negative) than other models. The confusion matrix and its visualization of these two models are shown in Fig 4 and Fig 5.

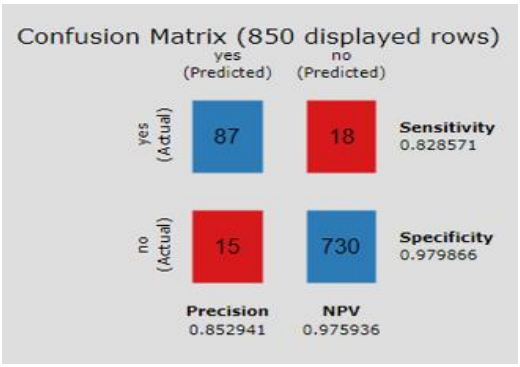

Fig 4. Confusion Matrix of GBT

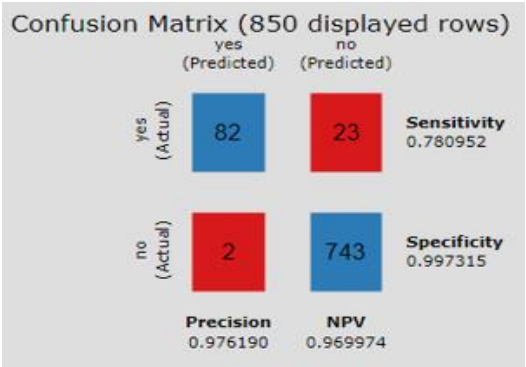

Fig 5. Confusion Matrix of RF

After examining the accuracy of all models, we choose the two models to build PMML (Predictive Model Markup Language) to predict the churn customers of the unlabeled dataset. In KNIME, each learner nodes produce the respective models. We used medium nodes such as "Gradient Boosted Tree to PMML" between the learner node and the PMML writer node. The "PMML writer" nodes that are externally produced PMML is used in phase 3 of Fig 2 to build two PMML models.

Fig 6 shows a typical application of the "PMML reader" node. The PMML file that is written to disk is passed to it. It is applied to unlabeled data that is read into KNIME. The unlabeled data is read by the CSV reader node, removed missing values, and filter the columns to correspond with the training data columns. JPMML (Java PMML) classifier can explore the result [12]. We filter the churn customers by using a row filter node and these records are reported to CRM.

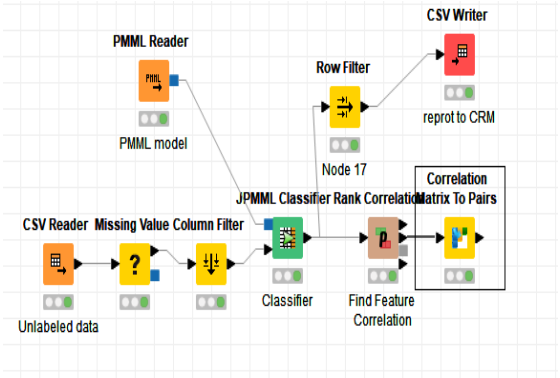

Fig 6. Example Workflow using PMML model

This evaluation obtained 93 churn customers from the GBT model and 78 churn customers from the RF model respectively when examined 750 unlabeled customers. The visualizations of this prediction result are presented with a pie chart in Fig 7 and Fig 8.

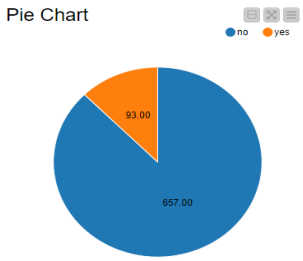

Fig 7. Churner and Non-churner Chart in GBT

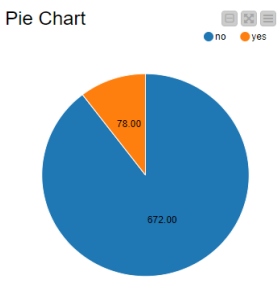

Fig 8. Churner and Non-churner Chart in RF

We used correlation analysis to determine the effectiveness of various features used in the prediction [13]. Correlation analysis is a statistical analysis of one variable to another. The data correlation of this analysis is the characteristics of each feature on how important for predicting the class label.

In this paper, each of the extracted features is compared with the final churn label in the dataset. This provides the similarity measure of each feature with the produced result and shows a direct dependence on the strength of the factor contributing to the churn. The matrix with the correlation coefficients for all pairs of data columns is shown in Fig 9. The blue color means full correlation (+1), white (0 = no correlation), red (-1 = full inverse correlation) [10]. The brighter the color, the more correlated between the two variables. Therefore, the lighter colors mean less correlation between them.

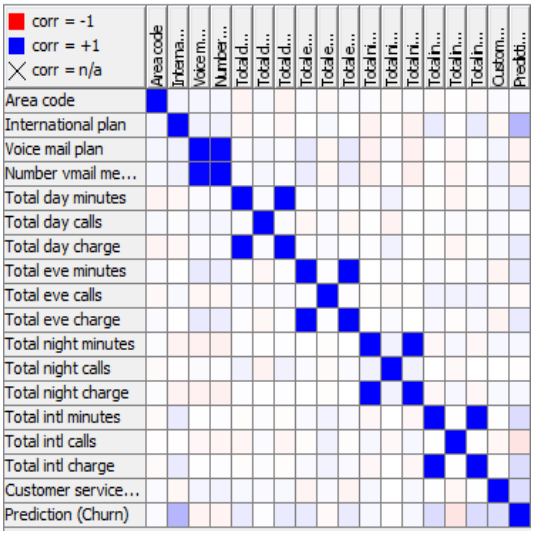

Fig 9. Correlation Matrix for Churn Parameters

The features that are important to prevent the churn rate are:

- International plan
- Total day minutes
- Total day charge
- Total intl minutes
- Total intl charge
- Customer service calls

These features have more correlation with the class label (prediction churn) because the intersection cells of each feature and class label have a brighter color of blue than other cells. This behavior of these features is in direct association with the customers who used a large number of day minutes, intl minutes, and expensed highly charged bills. Those customers that used more day minutes and intl minutes are more likely to be retained than that those of fewer usage minutes. Inversely, voice mail plan, voice mail message, and total intl calls features are light red in this matrix, the reason is the less use of these plans means the more chance to be churn. Thus, CRM would retain the customer churns carefully respect the above features.

VI. Conclusion

Nowadays, as many as the competitive market in telecom providers, churn is a very important role for CRM to retain valuable customers. Keep track of customer behavior among a large number of customer churn records is not easy using the traditional method. Therefore, machine learning models are applied to find out consumer behavior. In this work, the telecom data is trained with five predictive models and we selected Gradient Boosted Tree and Random Forest models because of the highest accuracy of 96.2 % and 96.89%. Then, we build PMML models and predict unlabeled data with these models. Using correlation analysis, the important features are extracted for CRM. Finally, we provide a powerful tool for organizations to determine which behavior of customers to focus upon for retaining and avoiding those customers lost to other providers.

References

[1] I. Ullah, B. Raza, A. K. Malaik, M. Imran, “A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector”. IEEE Access, Volume 7, 2019.

[2] Bart Baesens, “Analytics in a Big Data World” book. Published in 2014.

[3] A. Hammoudeh, M. Fraihat, M. Almomani, “Selective Ensemble Model for Telecom Churn Prediction”. Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), 2019.

[4] A. Alamsyah, N. Salma, “A Comparative Study of Employee Churn Prediction Model”. 4th International Conference on Science and Technology (ICST), Yogyakarta, Indonesia, 2018.

[5] M. Karanovic, M.Popocac, S. Sladojecic, M. Arsenovic, D. Stefanovic, “Telecommunication Services Churn Prediction Deep Learning Approach”. 26th Telecommunications forum TELFOR 2018, November 2018.

[6] K. Jang, J.Kim, B. Yu, "Vector-based Churn Prediction using Neural Networks in Mobile Games". 2019 IEEE International Conference on Big Data.

[7] Mr. A. Bhanagar, Dr. S. Srivastava, “Performance Analysis of Hoeffding and Logistic Algorithm for Churn Prediction in Telecom Sector”. 2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM).

[8] A. Amin, B. Shan, A.M. Khattak, T. Baker, "Just-in-time Customer Churn Prediction: With and Without Data Transformation”. 2018 IEEE Congress on Evolutionary Computation (CEC).

[9] Mr. A. Bhatnagar, Dr. S. Srivastava, “A Robust Model for Churn Prediction using Supervised Machine Learning”. 9th International Conference on Advanced Computing (IACC), 2019.

[10] www.knime.com.

[11] S. M. Basha, A. Khare, J. Gadipalli, “Training and Deploying Churn Prediction Model using Machine Learning Algorithms”. International Journal of Engineering Research in Computer Science and Engineering (IJERCSE), Vol 5 Issue 4, April 2018.

[12] D. Morent, K. Stathatos, “Comprehensive PMML Preprocessing in KNIME”. www.knime.com, 2020.

[13] S. Agrawal, A. Das, A. Gaikwad, “Customer Churn Prediction Modelling Based on Behavioural Pattern Analysis using Deep Learning”. International Conference on Smart Computing and Electronic Enterprise (ICSCEE 2018).