ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018

Analyzing Telecom Churn Trends Using ML Techniques

Athira V[#1], Jeena Thomas[*2]

[#]PG Scholar
St. Josephs College of Engineering and Technology, Palai
[1]athira24v@gmail.com

[*] Assistant Professor
St. Josephs College of Engineering and Technology, Palai
[2]jeena.thomas@sjcetpalai.ac.in

***Abstract*— Media transmission and in addition the network access suppliers assumes an essential part in the mobile communications and the data society. Telecom segment is consistently developing with its progression in highlights and innovation. Consumer loyalty is an imperative factor which choose the execution of firms. Customer churn prediction is an imperative test in CRM (Customer Relationship Management). Clients should be held to keep up the important clients. There are a few data mining procedures accessible like Decision Tree, Logistic Regression, Random Forest and Support Vector Machines. This paper gives an examination between various data mining methods in foreseeing churn and found that Gradient Boosting Classifier performs better that all others compared. Future Research Issues are additionally talked about.**

Key words: SVM, Decision Tree, Random Forest, Machine Learning, Churn Prediction

I. Introduction

Client Relationship Management is a vital disposition that should exist, it causes the organization to keep a cooperation with clients. This technique is more often focused in contemplating the historical backdrop of clients and make appropriate strategies good for clients in coopering with an organization to hold the client [20]. A few strategies like data mining, pattern recognition and correlation are utilized to examine client information, every one of these techniques discovered issues looked by clients.

As market rivalry builds client administration is an imperative methods for advantage for organizations. Diverse churn prediction models that exists does not perform well when managing enormous information [1]. Organizations give careful consideration to existing clients to maintain a strategic distance from client churn. Customer churn alludes to loss of clients who change from one supplier to its rival. Churning of a client can make incredible monetary misfortune an organization and even debilitate its reality. Churn prediction is a standout among the most imperative issues in customer relationship administration (CRM). Its point is to hold important clients to boost the benefit of an organization. To foresee whethera client will be a churner or non-churner, there are various information digging strategies connected for agitate forecast, for example, artificial neural networks, decision trees, and support vector machines [2].

Customer churn management is an imperative factor in service industry. There are distinctive approaches to deal with customer churn like data mining, machine learning and statistical models. Customer churn is a troublesome circumstance to deal with. It is hard to give a clarification like why a specific organization loss its clients. There can be various types of interior and in addition outer variables, social networking sites, celebrities [1] [3], companions can be a factor. Churn is an issue in telecom sector as it can influence itsexistence. They have to keep up the clients. In the aggressive business world a large portion of the endeavors stand in view of the benefit that originates from clients. Customer Relationship Management (CRM) dependably focuses on affirmed clients that are the most fertile source of information for decision making. That is, profoundly competitive associations have comprehended that holding existing and important clients is their core administrative strategy to get by in their businesses. However, holding clients is a troublesome assignment. Here comes the significance of churn management.

Churning of a customer will unfavorably influence the business, Customer churn can be viewed as clients who are proposing to move to a competing service provider. There are distinctive methods accessible to foresee whether a client will churn or not. There are different types of prediction available like churn prediction, insolvency prediction, and Fraud detection [5][6] .Telecom industry iso ne of there area field moving toward churn prediction. There are an immense number of wide band network available and the quantity of records is as yet expanding.

II. RELATED WORK

There are diverse strategies that deal with customer churn issue productively, yet they were just restricted to process little measure of information, with the utilization of cell phones and web there exist a need to deal with extensive measure of information. Trouble for churn management lies in choosing a systematic algorithm that can stay strong with enormous information in the business [11]. Several data mining methods and models are applied to predict churn of a customer. N.Kaur et.al [7] has given a hybrid approach on boosted tree for predicting churn, this new approach is an improvement of boosted tree, primarily focused on foreseeing the customer churn in an organization with a well defined model.

Castro et.al [4] proposed a frequency analysis approach in view of k-nearest neighbour machine learning algorithm for feature representation from login records for churn prediction, utilized for foreseeing churn of players in web based games. Chan et.al [8] proposed another calculation called DMEL (Data mining by Evolutionary Learning). It is a data mining algorithm to deal with classification problems, to predict churn under various churn rates with telecom subscriber information. Decision Tree, neural network and k-means [9] are chosen as a primary method to construct perspective models for telecom churn prediction. It gives more exact prediction. N. Lu et.al [10] proposed a churn prediction model utilizing boosting algorithm, which were mostly applied in banking industry. Boosted tree technique is utilized as a strong classifier algorithm using weak classifiers by performing certain iterations.

Hassouna et.al [12] has looked at decision tree and logistic regression model and found that decision tree performs superior to logistic regression models. A.Sharma et.al [13] proposed a neural system based way to deal with foresee customer churn in cellular network services, its outcome showed that the displayed approach performs better for customer churn prediction. J. Jadav et.al [5] [6] says data mining methods can be applied in understanding different issues in telecommunication sectors like churn prediction, insolvency prediction and fraud detection. M. Weiss et.al has given an answer for settle issues using data mining which incorporates into which initial step is tied in with understanding the information.

Li Shang et.al [15] has completed a preparatory report about applying data mining strategies to take care of customer churn issue in telecom sector, and for experimentation has utilized a data set from Taiwan mobile service company, and has investigated using logistic regression technique and decision tree, and has discovered that logistic regression has performed superior to decision tree. G. Toblerean [16] has predicted churners and non-churners utilizing SVM algorithm, and used a data set of twenty one attributes. A. Hudaib et.al [17] in order to upgrade the current models three hybrid models has been created for churn prediction. Clustering phase and prediction phase are the two stages that exist. In the first place customer information is filtered and afterward later prediction of customer churn is finished.

III. DATA MINING APPROACH

There are different models implemented using data mining technique using neural network, SVM and logistic regression. Customers are of different types like loyal customer, churn customer, voluntary churn and involuntary churn [7].Handling these customers to avoid churn is churn management. People who is satisfied with the services and products offered by a company are loyal customers, and those who are not satisfied shows an eagerness to churn known as churn customers. Voluntary churn occurs when customer itself cancels their relationship with the company, involuntary churn occurs when company itself cancels the relationship

A. Data Pre-Processing

Data in the real world is always noisy, First step that will be data munging or pre-processing. There can be errors in data, major task in pre-processing is data cleaning, data integration, data reduction and data discretion [2]. Data cleaning is required since data obtained will have some missing values,there is a task to fill the missing values, identify unwanted data,and resolve all these using necessary attributes. Data is usually combined from multiple sources as a single store, it is the integration process. Data obtained from different source will have different names, so necessary attributes should be selected properly. Care must be taken in integrating data from multiple source. Redundancy should be avoided to improve processing speed.After data integration it comes data transformation, it keeps the data in a small specified range.

B. Data Prediction/ Classification

After data is processed necessary algorithms are applied to the processing data. According the problems arise algorithm is selected. Supervised learning algorithms can be used like neural network, decision tree, these can be applied to find relationship between variables.

IV. METHODOLOGY

In this paper churn prediction is done using several existing data mining algorithms like Logistic Regression, Gradient boosting classifier, Decision Tree, Random Forest and SVM. All these algorithms are compared and found that Gradient Boosting Classifier algorithm performed better than all.

A. *Data Set*

Telecom churn dataset is used, it is an open source dataset which is freely available. Some of the features chosen from data set are given below.

TABLE I
ATTRIBUTES FOR CHURN PREDICTION

No:	**Features**
1	Account Length
2	Area Code
3	Number of voice mail messages
4	Total day calls
5	Total evening calls
6	Total night calls
7	Total Intl calls
8	Customer service calls

The dataset comes from the UCI machine learning repository. The dataset includes data related to telecom.It provides customer's information and includes 5000 records and 21 features. Among that some of the features are included in table 1. It gives customer information like his total day calls done, how many times one have called a customer service, all these information's help to understand the nature of a particular customer, and allows the company to take necessary decisions to retain their existing customer and to avoid churning of their customers.

B. *Logistic Regression*

Logistic regression [19] is a data mining technique used to predict occurrence probability of customer churn. It is based on a mathematically-oriented approach to analyze the variables affecting on others. Prediction is made by forming a set of equations connecting input values (i.e., affecting customer churn) with the output field (probability of churn). The equations (1), (2) and (3) give the mathematical formulas for a logistic regression model [19].

$$p(y = 1|x_1, ...x_n) = f(y) \quad (1)$$

$$f(y) = 1(1 + e - y) \quad (2)$$

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + ... + \beta nxn \quad (3)$$

Here y is the target variable for each individual j, 0 is constant 1 is the weight given to the specific variable associated with each customer j, X1...xn is the predictor variables.

A. *Random Forest*

One of the most popular ensemble classier for decision trees is known as the Random Forests algorithm. The algorithm works by splitting the dataset into random subsets of samples and subsequently generating decision trees on each subset. During the prediction phase each tree is allowed to report its predictions and the majority prediction is the one returned by the model .Random forest are ensemble learning method for classification, regression and many other tasks. Random forest can be used for both classification and regression type of problems. Random Forest pseudocode:

```
1) Randomly select k features from total m features.
Where k! = m

2) Among the k features, calculate the node d using
the best split point

3) Split the node into daughter nodes using the best
split.

 4) Repeat 1 to 3 steps until l number of nodes has
been reached

5) Build forest by repeating steps 1 to 4 for n
number times to create n number of trees.
```

B. *SVM*

These are the supervised models with associated learning algorithms that can be used to analyse data for classification and regression analysis. SVM model is a representation of the examples as points in space. Support Vector Machines only require two parameters to be chosen in order for them to generate predictions. The kernel parameter and slack variable C, The model generated by SVMs is always optimal and global. SVMs perform very well in the application of churn prediction even on realistic, noisy datasets. SVMs take significantly more time to train than Logistic Regression and Random Forests, it is a drawback of SVM. Version of SVM for regression is given

$$minimize 1/2||w||^2 \tag{4}$$

subject to

$$y_i - (w, x_i) <= \epsilon \tag{5}$$

$$(w, x_i) + b - y_i <= \epsilon \tag{6}$$

C. Gradient Boosting Classifier

Gradient boosting is a machine learning technique for regression and classification problems, it gives a prediction model in the form of an ensemble weak prediction models like decision tree. GB builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage n classes of regression trees are fit on the negative gradient of the binomial or multinomial deviance loss function. Binary classification is a special case where only a single regression tree is induced. At each stage of gradient boosting, it assumes that there is some imperfect model Fm, and gradient boosting improves Fm. by constructing a new model.

$$F_m + 1(x) = F_m(x) + h(x) = y \tag{7}$$

D. K – Nearest Neighbour

The k-nearest neighbour is a data mining algorithm mostly used for classification. The algorithm can also be used for estimation and prediction. The k in the model determines the number of variables included in the neighbourhood. If there are continuous response variables, the nearest neighbour value given to each variables response (yi) is determined by equation below

$$y_i = 1/k \sum_{x_j \epsilon N x_i} \tag{8}$$

Where x corresponds to theneighbourhood of xi, N (xi) and k is a fixed constant, Nearest-neighbour has two methods: the distance function and the cardinality k.

V. PERFORMANCE EVALUATION

Different ensemble methods are evaluated and Gradient Boosting Classifier performs better compared with other methods. Confusion matrix shows accuracy. Here it is seen that random forest performs better than SVM, as SVM is a lazy evaluation technique.

TABLE II
ACCURACY OBTAINED

Algorithm Used	Accuracy
Logistic Regression	0.857
Gradient Boosting Classifier	0.915
Support Vector Machine	0.903
Random Forest	0.908
K – Nearest Neighbour	0.809

Table II has given accuracy obtained after applying necessary algorithms in the taken dataset and has found that gradient boosting classifier performs better than any other method. It is difficult to understand the vector plain of SVM. Most of the datasets will have its own randomness there will be some noise in that and then an ensemble method will perform better.

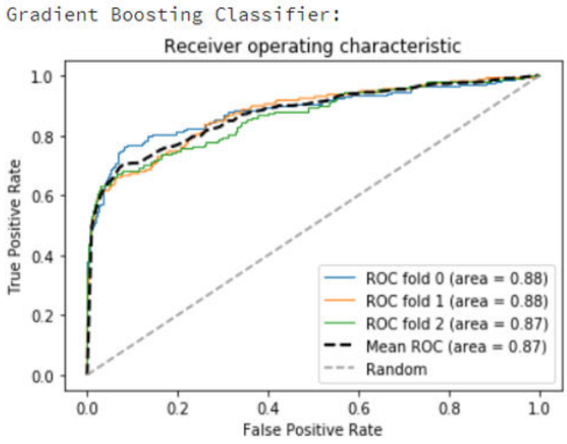

Fig. 1 Gradient Boosting Classifier

A perfect system will be like Receiver Operating Characteristic (ROC) will be one fold indicates the fold of data, if we have hundreds of data observations then we divide the data into folds. ROC curve shows the relationship between the true positive rate and the false positive rate. It gives the relation between churners ratio and non- churners ratio, which gives the correct prediction of people who is churned and not churned.In Gradient Boosting we can find that the variance is minimal when compared to Random forest. We can find Random forest and gradient boosting will have competition because both are ensemble methods.

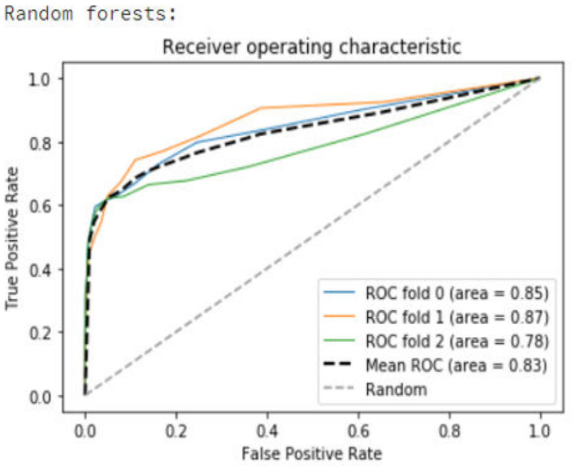

Fig. 2 Random Forest

Figure 1 and 2 has given the ROC characteristics of random forest and gradient boosting classifier, and has found gradient boosting has the better rate compared to random forest. A model is said to be best performing one when its ROC curve passes through or is close to (0, 1). Diagonal line divides the ROC space into two parts, ROC curve passing near this line correspond to random guessing classifier. Area under the ROC curve (AUC) is used as a performance metric, whose values ranges from 0.0 to 1.0. Models perform better when they have greater AUC.

VI. CONCLUSION

Businesses in the consumer market and in all enterprise sectors have to deal with churn. Sometimes churn is excessive and influences policy decisions. Sophisticated handling of churn is a sign of a mature industry. The classic example is the telecommunications industry where subscribers are known to frequently switch from one provider to another. This voluntary churn is a prime concern. This paper has compared different machine learning algorithms in predicting churn. From this it is observed that Gradient Boosting Classifier performs better than Random Forest and other ensemble methods. Selecting the right attribute and proper values may result in more accurate result. Future research direction may include deep learning models to predict churn.

REFERENCES

[1] Wenjie Bi, “A Bigdata clustering algorithm for mitigating the risk of customer churn,” IEEE Transactions on Industrial Informatics, 1551-3203 March 2015.

[2] Yu-Hsin Lu, “Data mining techniques in customer churn prediction,” Recent Patents on Computer Science 2010, vol.3, pp. 28-32.

[3] B. Brian,” Customer churn prediction in telecommunications,” Expert Syst. Appl, vol. 39, no. 1, pp. 1414-1425, Jan 2012.

[4] E.G. Castro, “Churn predictionin online games using players login records: a frequency analysis approach,” IEEE Trans. Compute. Intell, AI Games, vol. 7, no. 3, pp. 255-265, Sep 2015.

[5] S. Babu, “A Review on customer churn prediction in telecom using datamining techniques,” International journal of scientific engineering and research, vol.4, Issue 1, Jan 2016.

[6] T.Pawar,” Churn prediction in telecommunication using datamining technology,” vol. 2, no.2, Feb 2011.

[7] Navneet Kaur,” churn prediction- A review,” vol. 4, Issue 1, January 2016.

[8] Y.Yin,,” A novel evolutionary data mining algorithm with applications to churn prediction,” IEEE Transactions on Evolutionary Computing, vol. 7, no.6, pp. 532-545, Dec 2003.

[9] S.Y. Hung,” Applying data mining to telecom churn management,” vol.31, no.3, pp.515-524, Oct 2006.

[10] N.Lu,” A customer churn prediction model in telecom industry using boosting,” IEEE Transactions, vol.10, no.2, pp- 1659-1665, May 2014.

[11] C.Y. Zhang,” Data-intensive applications, challenges, techniques and technologies: A survey on big data,” vol. 275, pp.314-347, Aug. 2014.

[12] Ali Tarhini,” Customer churn in mobile markets: A comparison of techniques, vol. 8, no.6, 2018.

[13] A Sharma,” A neural network based approach for predicting customer churn in cellular network services,” International journal of computer application, vol. 27, no. 11. Aug 2011.

[14] M.Weiss,” Data mining in telecommunications,” Department of computer science.

[15] Li- Shang,” Knowledge discovery on customer churn prediction,” Nov 2006.

[16] Gavril Toblerean,” Churn prediction in the telecommunication sector using support vector machines,” May 2013.

[17] Amjad Hudaib,” Hybrid Data Mining Models for Predicting Customer Churn,” vol. 8, 2015.

[18] Pruti. K. Dalvi,” Analysis of customer churn prediction in telecom industry using decision tree and logistic regression”, IEEE 2016.

[19] Ali Tarhini,” Customer Churn in Mobile Markets: A Comparison of Techniques,” vol. 8, no. 6; 2015.

[20] Management Tools-Customer Relationship Management Bain and Company, www.bain.com, Nov 2015.