PREDICTING CUSTOMER CHURN IN TELECOM SECTOR USING RANDOM FOREST CLASSIFICATION

[1]Saurabh Singour, [2]Jitendra Agarwal, [3]Sanjeev Sharma, [4]Shikha Agrawal
[1,2,3,4]Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal M.P. India

Abstract-Customer churn analysis and prediction in telecom sector is an issue now a days because it's very important for telecommunication industries to analyze behaviors of various customer to predict which customers are about to leave the subscription from telecom company. So data mining techniques and algorithm plays an important role for companies in today's commercial conditions because gaining a new customer's cost is more than retaining the existing ones. In this paper we can focuses on data mining algorithm for predicting customer churn, In this we can build the classification model using random forest techniques and model evaluation measures are computed and compare with logistic regression model.

Keywords- *Churn prediction;data mining, telecom system;Customer retention;classification system; random forest;logistic regression.*

I. INTRODUCTION

Studies revealed that gaining new customers is costlier than keeping existing customers happy and loyal in today's competitive conditions, and that an average company loses 10 to 30 percent of customers annually. Many companies, being aware of this fact, are engaged in satisfying and retaining the customers. Especially in the subscription oriented industries, such as telecommunications, banking, insurance, and in the fields of customer relationship management, etc., companies working with numerous customers, the revenues of the companies are provided by the payments made by these customers periodically. It is very important to be able to keep customers satisfied in order to be able to sustain this revenue with the least expenditure cost.

The objectives of this study are:

• Reviewing the relevant studies about churn analysis on telecommunications industry presented in the last five years, particularly in the last two years, and introducing these up-to-date studies in the literature,

• Determining the data mining methods frequently used in churn implementations,

• Shedding a light on methods that can be used in further studies.

Data Mining and Customer Churn Analysis

In today's technological conditions, new data are being produced by different sources in many sectors. However, it is not possible to extract the useful information hidden in these data sets, unless they are processed properly. In order to find out these hidden information, various analyses should be performed using data mining, which consists of numerous methods.

The Churn Analysis aims to predict customers who are going to stop using a product or service among the customers. And, the customer churn analysis is a data mining based work that will extract these possibilities. Today's competitive conditions led to numerous companies selling the same product at quite a similar service and product quality.

With the Churn Analysis, it is possible to precisely predict the customers who are going to stop using services or products by assigning a probability to each

customer. This analysis can be performed according to customer segments and amount of loss (monetary equivalent). Following these analyses, communication with the customers can be improved in order to persuade the customers and increase customer loyalty. Effective marketing campaigns for target customers can be created by calculating the churn rate or customer attrition. In this way, profitability can be increased significantly or the possible damage due to customer loss can be reduced at the same rate . For example, if a service provider which has a total of 2 million subscribers, gains 750.000 new subscribers and losts 275.000 customers; churn rate is calculated as 10%. The customer churn rate has a significant affect on the financial market value of the company. So most of the companies keep an eye on the value of the customer at monthly or quarterly periods.

II. LITERATURE REVIEW & OBJECTIVES

According to the paper [1] , From the beginning of the data mining which is used to discover new knowledges from the databases can helping various problems and helps the business for their solutions. Telecom companies improve their revenue by retaining their customers Customer churn in telecom sector is to leave a one subscription and join the other subscription In these paper they predicting the customer churn by using various R packages and they created a classification model and they train by giving him a dataset and after training they can classify the records into churn or non churn and then they visualize the result with the help to visualization techniques. In this they are using logistic regression model and these model first train on training data after that they can test the model on test data to compute the performance measure of the classification model so we can get the various parameters like true positive rate, false positive rate and accuracy.

According to [2] , Telecom Customer churn prediction is a cost sensitive classification problem. Most of studies regard it as a general classification problem use traditional methods, that the two types of misclassification cost are equal. And, in aspect of cost sensitive classification, there are some researches focused on static cost sensitive situation. In fact, customer value of each customer is different, so misclassification cost of each sample is different. For this problem, we propose the partition cost-sensitive CART model in this paper. According to the experiment based on the real data, it is showed that the method not only obtains a good classification performance, but also reduces the total misclassification costs effectively.

According to paper [5] Customer churn plays an important role in customer relationship management (CRM), and they are using various machine learning algorithm to predict customer churn and they found ensemble learning is an best to predict customer churn, but there exist still a lot of problems like how they choose the method of integration and how to choose the strategy, which makes the final ensemble classifier. On the other hand, there is no good classifier, so its also a main problem to chosen which classification algorithm is best for which situation. So we can consider various aspect like vertical and horizontal contrast to find the best classifier to predict the customer churn in telecom sector.

A. Objectives

Telecom industry market camps shows that their is more than 20-40 % loss of customer in the industries. So it is very important to know about customer behaviour to find why the customer is left their subscription and join the other company subscription. So reducing customer churn in telecom industries in a problem which is facing by various companies.

There are various reason for customer churn analysis:

1) Maximizing the profit by reducing the market
2) By predicting customer churn they can reduce churn.
3) Identify those customers who will generates more profits
4) Finding customer behaviour
5) Cost because adding a new customer is costlier than retaining a customer.
6) For making cost effective marketing strategies

III. PROBLEM DEFINITION & MOTIVATION

In a business, client attrition merely refers to the purchasers exploit one service to a different. customer

churn is kind of problem in which a customer shift from one service to a other service. Machine learning techniques in which churn prediction could be a supervised and it follows as : the goal is to predict the customer churn in telecom over that horizon, given the info related to every subscriber within the network. The churn prediction involves three phases, namely, i) the training phase, ii) testing phase, iii) prediction phase. The input telecom datasets includes the information on past necessitate every mobile subscriber, along with all personal and business data that's maintained by the service supplier. And for the training section, labels or attributes are provided within the type of an inventory of churners. when the model is trained, the model should be able to predict the list of churners from the important dataset that doesn't embody any churn label.

A. Motivation

There are several issues facing by telecom for data mining.

1) There are billions of records contains the telecom datasets.

2) The raw data is directly not suitable for data mining.

The raw data which is extracted from telecom sector contains redundancy, missing records so data summarization is time taking tasks from data mining. This paper focus on identifying the customer behavior through which we can identify those customers who may or may not leave the company subscription. The analysis of telecom datasets will help us to understand customer behaviour and satisfaction with the company services.

IV. PROPOSED WORK

In these we can proposed an machine learning technique through which we can developed a classification model based on random forest algorithm to predict customer churn. These classification model are build by using R programming which is an widely used software for machine learning and data mining.

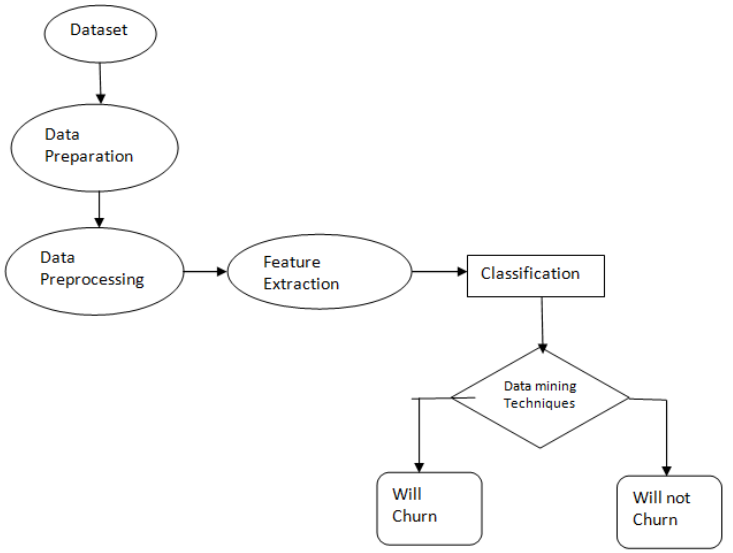
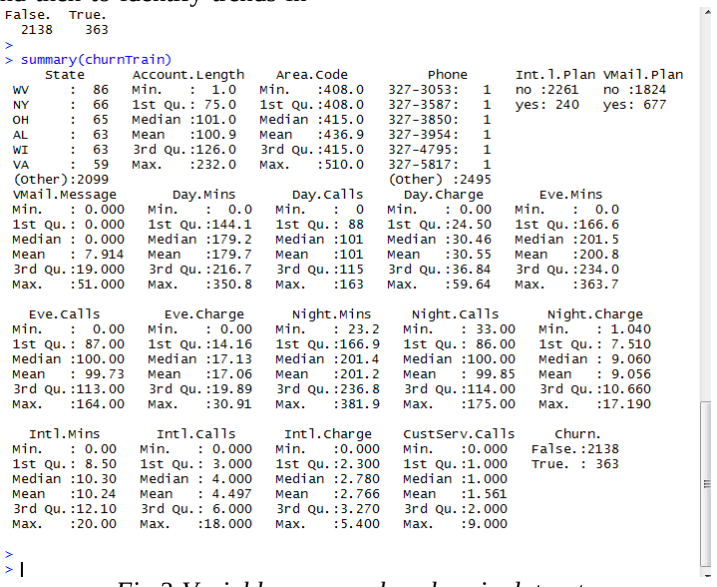


Fig.1:Churn Prediction Framework

In these we can take a telecom dataset and first we can prepare the dataset because we cannot apply the dataset directly to the classification model so we can first prepare the dataset with respect to model by giving proper attributes names. After preparing we can preprocess the dataset through which we can remove redundancy or missing records which can affects the learning of classification models. After preprocessing we can giving the training dataset to the model to learn and from these dataset the model can extracted some features and based on these feature it will classify the test datasets into churners or non churners.

V. EXPERIMENT & RESULT ANALYSIS

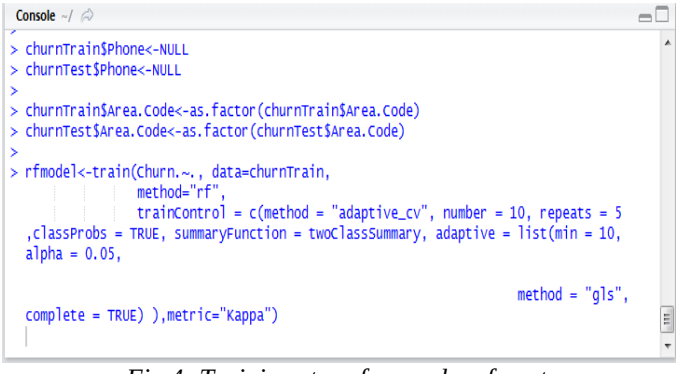
All the result analysis which we can perform by using an intel i5 CPU with 2.30 GHz processor and 4 GB of RAM running Windows. Then we can install r and rstudio for data mining and then to identify trends in customer churn at a telecom company. The dataset given to us contains 3,333 observations and 23 variables extracted from a data warehouse. These variables are shown in figure 3.



```
False.  True.
  2138    363
>
> summary(churnTrain)
     State         Account.Length    Area.Code          Phone      Int.l.Plan VMail.Plan
 WV      :  86   Min.   :  1.0   Min.   :408.0   327-3053:   1   no :2261   no :1824
 NY      :  66   1st Qu.: 75.0   1st Qu.:408.0   327-3587:   1   yes: 240   yes: 677
 OH      :  65   Median :101.0   Median :415.0   327-3850:   1
 AL      :  63   Mean   :100.9   Mean   :436.9   327-3954:   1
 WI      :  63   3rd Qu.:126.0   3rd Qu.:415.0   327-4795:   1
 VA      :  59   Max.   :232.0   Max.   :510.0   327-5817:   1
 (Other):2099                                    (Other) :2495
 VMail.Message       Day.Mins       Day.Calls     Day.Charge       Eve.Mins
 Min.   : 0.000   Min.   :  0.0   Min.   :  0   Min.   : 0.00   Min.   :  0.0
 1st Qu.: 0.000   1st Qu.:144.1   1st Qu.: 88   1st Qu.:24.50   1st Qu.:166.6
 Median : 0.000   Median :179.2   Median :101   Median :30.46   Median :201.5
 Mean   : 7.914   Mean   :179.7   Mean   :101   Mean   :30.55   Mean   :200.8
 3rd Qu.:19.000   3rd Qu.:216.7   3rd Qu.:115   3rd Qu.:36.84   3rd Qu.:234.0
 Max.   :51.000   Max.   :350.8   Max.   :163   Max.   :59.64   Max.   :363.7

   Eve.Calls        Eve.Charge      Night.Mins     Night.Calls      Night.Charge
 Min.   :  0.00   Min.   : 0.00   Min.   : 23.2   Min.   : 33.00   Min.   : 1.040
 1st Qu.: 87.00   1st Qu.:14.16   1st Qu.:166.9   1st Qu.: 86.00   1st Qu.: 7.510
 Median :100.00   Median :17.13   Median :201.4   Median :100.00   Median : 9.060
 Mean   : 99.73   Mean   :17.06   Mean   :201.2   Mean   : 99.85   Mean   : 9.056
 3rd Qu.:113.00   3rd Qu.:19.89   3rd Qu.:236.8   3rd Qu.:114.00   3rd Qu.:10.660
 Max.   :164.00   Max.   :30.91   Max.   :381.9   Max.   :175.00   Max.   :17.190

   Intl.Mins       Intl.Calls     Intl.Charge    CustServ.Calls     Churn.
 Min.   : 0.00   Min.   : 0.000   Min.   :0.000   Min.   :0.000   False.:2138
 1st Qu.: 8.50   1st Qu.: 3.000   1st Qu.:2.300   1st Qu.:1.000   True. : 363
 Median :10.30   Median : 4.000   Median :2.780   Median :1.000
 Mean   :10.24   Mean   : 4.497   Mean   :2.766   Mean   :1.561
 3rd Qu.:12.10   3rd Qu.: 6.000   3rd Qu.:3.270   3rd Qu.:2.000
 Max.   :20.00   Max.   :18.000   Max.   :5.400   Max.   :9.000

>
>
```

Fig.3:Variables or sample values in datasets

After that we can build a classification model by using random forest, and we can train the model by applying telecom dataset to these model. We can provide 75% data for training and 25% for testing , figure 4 shows the training steps for random forest model.



```
Console ~/
> churnTrain$Phone<-NULL
> churnTest$Phone<-NULL
>
> churnTrain$Area.Code<-as.factor(churnTrain$Area.Code)
> churnTest$Area.Code<-as.factor(churnTest$Area.Code)
>
> rfmodel<-train(Churn.~., data=churnTrain,
                 method="rf",
                 trainControl = c(method = "adaptive_cv", number = 10, repeats = 5
,classProbs = TRUE, summaryFunction = twoClassSummary, adaptive = list(min = 10,
alpha = 0.05,

                                                          method = "gls",
complete = TRUE) ),metric="Kappa")
```

Fig.4: Training steps for random forest

After training we can test the model by applying remaining 25% of data and find the evaluation measure which is shown in figure 5.

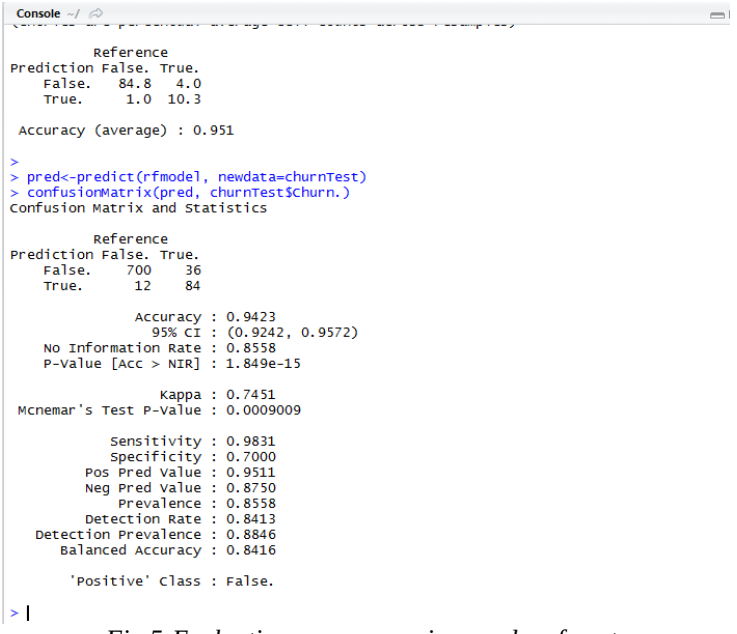


Fig.5:Evaluation measures using random forest.

CLASSIFICATION ALGORITHM	ACCURACY
LOGISTIC REGRESSION	80.02%
RANDOM FOREST	95.10%

Table 1. Classification measures

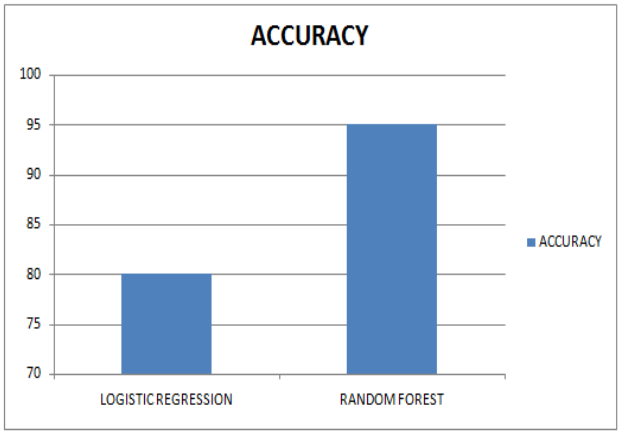


Fig.6:Classification evaluation measures

Random Forests grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. So the learning of this model will take time, but when the model gets trained its produces very accurate result as compared to other models figure 6 clearly shows the difference in accuracy with logistic regression.

VI. CONCLUSION

For retaining the existing customer in telecom sector we need a reason why the customer is churn by analyzing the customer behaviour. In this paper, it is observed that random forest model better in the prediction of churn because it has more accuracy than logistic regression classification model and it is also easy to construct.

REFERENCES

[1]. Helen Treasa Sebastian* and Rupali Wagh, " Churn Analysis in Telecommunication using Logistic Regression " in OJCST, March 2017, Vol. 10,: ISSN: 0974-6471, Pno. 207-212.

[2]. Chuanqi Wang, Ruiqi Li, Peng Wang, Zonghai Chen, "Partition cost-sensitive CART based on customer value for Telecom customer churn prediction" in Proceedings of the 36th Chinese Control Conference 2017 IEEE.

[3]. Guo-en Xia, Hui Wang, Yilin Jiang, "Application of Customer Churn Prediction Based on Weighted Selective Ensembles" in IEEE 2016.

[4]. Rahul J. Jadhav, Usharani T. Pawar, "Churn Prediction in Telecommunication Using Data Mining Technology", in *(IJACSA), Vol. 2, No.2, February 2011*

[5]. Kiran Dahiya, Surbhi Bhatia, "Customer Churn Analysis in Telecom Industry" in IEEE 2015, 978-1-4673-7231-2/15

[6]. N.Kamalraj, A.Malathi' " A Survey on Churn Prediction Techniques in Communication Sector" in *IJCA Volume 64– No.5, February 2013*

[7]. Kiran Dahiya,KanikaTalwar, "Customer Churn Prediction in Telecommunication Industries using Data Mining Techniques- A Review" in IJARCSSE, Volume 5, Issue 4, 2015.

[8]. R Data: http://cran.r-project.org/

[9]. Data Mining in the Telecommunications Industryl, Gary M. Weiss, Fordham University, USA.

[10]. Manjit Kaur et al., 2013.Data Mining as a tool to Predict the Churn Behaviour among Indian bank customers, IJRITCC, Volume: 1 Issue: 9

[11]. R. Khare, D. Kaloya, C. K. Choudhary, and G. Gupta, "Employee attrition risk assessment using logistic regression analysis,".

[12]. Praveen et al., Churn Prediction in Telecom Industry Using R, in (IJETR) ISSN: 2321-0869, Volume-3, Issue-5, May 2015

[13]. J. Burez and D. Van den Poel, "Handling class imbalance in customer churn prediction," *Expert Systems with Applications*, vol. 36, no. 3, 2009.

[14]. Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.