Analysis of Telecom Customer Churn Prediction by Building Decision Tree

Chandana S
4AI15CS029
Department of CS & E Adhichunchanagiri Institute of Technology, Chikmagalur

Vineetha G
4AI15CS123
Department of CS & E Adhichunchanagiri Institute of Technology, Chikmagalur

Varun E
Assistant Professor Department of CS & E
Adhichunchanagiri Institute of Technology,

Chikmagalur

Dr. Pushpa Ravikumar
Professor & Head Department CS &E
Adhichunchanagiri Institute of Technology,
Chikmagalur

Abstract:- **Telecommunication has become an important for business, enabling companies to communicate effectively with its customers and allowing high standards of customer service. Due to the expansion of telecommunication market day by day and increased competition has resulted in huge loss of revenue as well loss of customers. The process of one customer leaving one telecom company and joining another telecom company is called as "Churn".**

In this paper we developed a prediction model for telecom customer churn. It represents large dataset in the form of graphs which helps to depict the outcome in the form of various data visualization. Churn is a very important area in which the telecom domain can make or lose their customers hence investing greater time to make predictions which in turn helps to make necessary business conclusions. Churn reduction can be achieved effectively by analysing the past history of the potential customer systematically.

Keywords- Telecommunication, Churn, Churn Prediction, Churn Management, Churn rate, Data mining.

I. INTRODUCTION

The mobile communication has become the dominant medium for effective communication all over the world. In numerous countries, especially in developed, the telecom market is saturated to the extent that each new customer must be won over from the competitors. Many public policies and standardised procedures of mobile communication allow the customer to easily switch over from one carrier to another. So, instead of winning a new customer it is far better to retain the old customers in the same network. Hence, the telecom carriers have now shifted their focus from customer acquisition to customer retention.

Churn in terms of telecom industries refers to the customer leaving the current company and moving towards the telecom company. Managing of customer to remain in the particular telecom company is intact a difficult task. Customer churn is a notorious problem for most of the industries, affecting the revenues and standards of a company, subsequently resulting in difficulty for acquiring of new customers. In the customer oriented telecommunication cycle, churn refers to the decision made by the customer about ending up the business relationship with a particular telecommunication company due to prevalence of inconvenience over a long duration.

Churn may also be referred as loss of clients or customers, who are intending to move their custom to a competing service provider. In order to manage customer churn more effectively, a company must build an accurate and more effective churn prediction technique. To keep up in the competition and to acquire as many customers, most of the telecom service providers invest huge amount of revenue to expand their business in the beginning. Therefore, it has become important for the telecom operators to earn back the amount they invested along with at least the minimum profit within a very short period of time.

The Decision Trees, Nearest Neighbour, and Artificial Neural Networks are sum of the churn analysis techniques which perform two key tasks such as predicting whether a particular customer will churn and reasons for that particular customer to churn. These techniques address only percentage of churn, but they fail to identify the exact number of churners. The problem confronting wireless telecommunications management is that it is very difficult to determine which subscribers leave the company and why. It is therefore more difficult to predict which customers are likely to leave the company, and devise cost effective incentives that will convince likely churners to easy.

II. LITERATURE SURVEY

A. Churn Prediction

In the past, churn has been identified as an issue of concern across most industry sectors. In its most general sense it refers to the rate of loss of customers from a company's customer base[2]. The churning out of the customers from the emerging business space like telecom and broadcast providers leads to the loss of revenue. These companies aim at identifying the risk of churn in its early stages, as it is usually much cheaper to retain a customer than to try to win that customer back. If this risk can be accurately predicted, marketing departments can target customers efficiently with targeted incentives to prevent them from leaving. Generally the churn refers to the defection of a customer. In the telecom industry, a subscriber is to have churned when he leaves one network and moves towards the other telecom networks. Churn rate is defined as the total number of subscribers who leave the service in the period divided by average total customers in the period. The churn rate of a telecom company is a key measure of risk and loss of revenue in the telecom industry and it should be quoted in the company annual report[2].

B. Types of Churn:

Telecom churn can be mainly classified in two type's namely voluntary churn and involuntary churn. Voluntary churn also referred to as Active churn occurs when customer initiates service termination. It is purely based on the decision of the customer to unsubscribe from the existing service providers. These are further classified as deliberate and incidental churn. The reasons for the former include un satisfaction with service quality, high costs, privacy concerns, no fault resolution etc. While the latter happens due to changes in circumstances that prevent customer from further continuing with the service e.g. financial problems, change in geographical location of the customer. Involuntary churn means the company suspends the customer's service and no longer want them as subscribers of their company. This is usually because of non-payment, service abuse, fraud or non-use of service over longer duration[1].

Churn can be reduced by knowing the past history of the customers so that they can improve their services to fulfil the demands of the customer. Large amount of information about the customers such as billings, total calls, customer service centre calls and network data are maintained in a telecom companies. The availability of enormous amount of data arises the scope of using data mining techniques in the telecom database. The telecom dataset helps the service providers to easily analyse the data about the customer in different perspectives to predict and reduce the churning.

Factors Influencing the Customer churn are,

- Weak customer marketing.
- Poor customer service.
- Ineffective relationship building.
- Bad unboarding.
- Competitive driven churn.
- Customer did not achieve their deserved outcome.

C. Data Set Used: Telecom data set attributes

Table 1: Telecom Dataset Attributes

Sl.no	Attributes
1	Customer
2	Age
3	Area
4	Phone
5	International plan
6	Day minutes
7	Day calls
8	Day charges
9	Evening minutes
10	Evening calls
11	Evening charges
12	Night minutes
13	Night calls
14	Night charges
15	International minutes
16	International calls
17	International charges
18	Customer service calls
19	Churn
20	Occupation
21	Total calls

D. Data Mining Techniques

Data mining is a process of examining of already existing data from a large data set to create a new information. Also, data mining can be defined as the process of reducing, analysing the patterns, predicting the useful and required information from a large database and transform it into an understandable structure for future use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interesting metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. One of the most commonly used analysis step in the Data mining is the "Knowledge discovery in databases" process, or KDD. And the common data mining technique is Classification, its aim is to classify unknown cases based on the set of known examples into one possible classes. Here, we use Classification technique for the telecom dataset which helps telecom operators to predict whether the particular customer will churn or not.

III. METHODOLOGY

To find the answer for who and why is likely to churn, an effective classification of customer is much needed. Therefore, churn prediction deals with the identification of the customers who are likely to churn in near future.

The figure 1 illustrates the process involved in identifying and classifying the customers as churners and non-churners based on different perspectives.

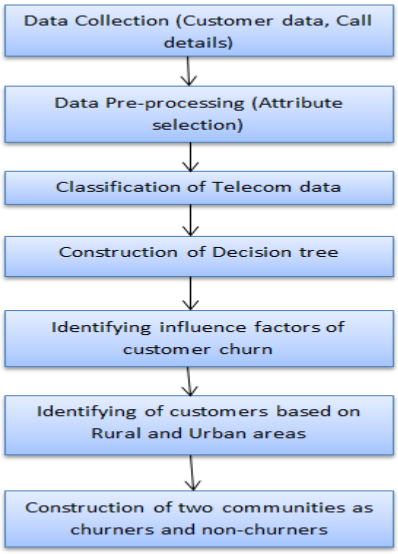

Data Collection (Customer data, Call details)
Data Pre-processing (Attribute selection)
Classification of Telecom data
Construction of Decision tree
Identifying influence factors of customer churn
Identifying of customers based on Rural and Urban areas
Construction of two communities as churners and non-churners

Figure 1: The process diagram of Churn Prediction

A. Churn prediction decision tree

Decision trees are the most commonly used tool for prediction and classification of future events. Each node in a decision tree is a test condition and branching is based on the value of the attribute being tested. The label of the leaf node (churner or non-churner) is assigned to the customer under evaluation.

The figure 2 illustrates a simplified churn prediction decision tree for a telecommunication Sector,

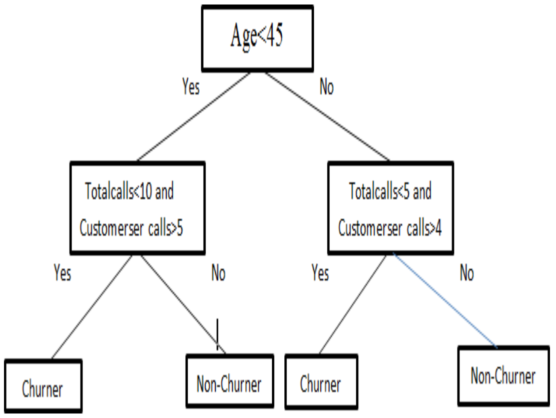

Age<45
Yes
No
Totalcalls<10 and Customerser calls>5
Totalcalls<5 and Customerser calls>4
Yes
No
Yes
No
Churner
Non-Churner
Churner
Non-Churner

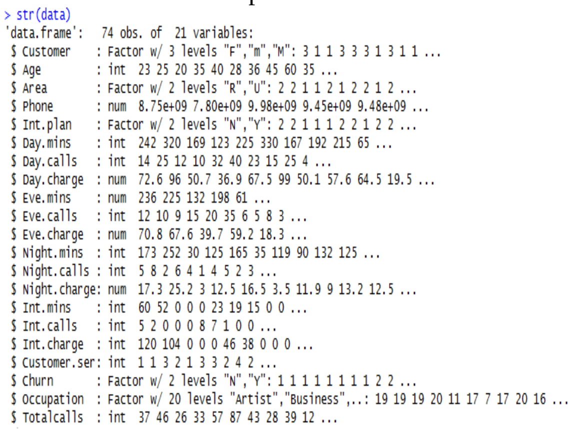
Figure 2: A Simplified Churn Prediction decision tree

Decision trees are more often criticized that they are not suitable for addressing complex and non- linear relationships between the attributes. But many researches have documented that accuracy of decision tree is high.

IV. EXPERIMENTAL RESULTS AND OBSERVATIONS

A. Description of the complete data set

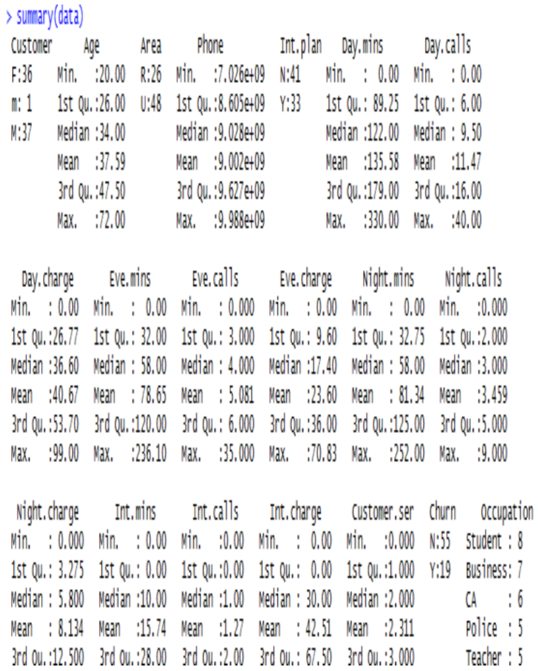
Table 2: Description of telecom data set



```
> str(data)
'data.frame':   74 obs. of  21 variables:
 $ Customer    : Factor w/ 3 levels "F","m","M": 3 1 1 3 3 3 1 3 1 1 ...
 $ Age         : int  23 25 20 35 40 28 36 45 60 35 ...
 $ Area        : Factor w/ 2 levels "R","U": 2 2 1 1 2 1 2 2 1 2 ...
 $ Phone       : num  8.75e+09 7.80e+09 9.98e+09 9.45e+09 9.48e+09 ...
 $ Int.plan    : Factor w/ 2 levels "N","Y": 2 2 1 1 1 2 2 1 2 2 ...
 $ Day.mins    : int  242 320 169 123 225 330 167 192 215 65 ...
 $ Day.calls   : int  14 25 12 10 32 40 23 15 25 4 ...
 $ Day.charge  : num  72.6 96 50.7 36.9 67.5 99 50.1 57.6 64.5 19.5 ...
 $ Eve.mins    : num  236 225 132 198 61 ...
 $ Eve.calls   : int  12 10 9 15 20 35 6 5 8 3 ...
 $ Eve.charge  : num  70.8 67.6 39.7 59.2 18.3 ...
 $ Night.mins  : int  173 252 30 125 165 35 119 90 132 125 ...
 $ Night.calls : int  5 8 2 6 4 1 4 5 2 3 ...
 $ Night.charge: num  17.3 25.2 3 12.5 16.5 3.5 11.9 9 13.2 12.5 ...
 $ Int.mins    : int  60 52 0 0 0 23 19 15 0 0 ...
 $ Int.calls   : int  5 2 0 0 0 8 7 1 0 0 ...
 $ Int.charge  : int  120 104 0 0 0 46 38 0 0 0 ...
 $ Customer.ser: int  1 1 3 2 1 3 3 2 4 2 ...
 $ Churn       : Factor w/ 2 levels "N","Y": 1 1 1 1 1 1 1 1 2 2 ...
 $ Occupation  : Factor w/ 20 levels "Artist","Business",..: 19 19 19 20 11 17 7 17 20 16 ...
 $ Totalcalls  : int  37 46 26 33 57 87 43 28 39 12 ...
```

B. *Summary of data set*

Table 3: Summary of data set



```
> summary(data)
Customer      Age         Area        Phone          Int.plan   Day.mins        Day.calls
F:36    Min.   :20.00   R:26   Min.   :7.026e+09   N:41    Min.   :  0.00   Min.   : 0.00
m: 1    1st Qu.:26.00   U:48   1st Qu.:8.605e+09   Y:33    1st Qu.: 89.25   1st Qu.: 6.00
M:37    Median :34.00          Median :9.028e+09           Median :122.00   Median : 9.50
        Mean   :37.59          Mean   :9.002e+09           Mean   :135.58   Mean   :11.47
        3rd Qu.:47.50          3rd Qu.:9.627e+09           3rd Qu.:179.00   3rd Qu.:16.00
        Max.   :72.00          Max.   :9.988e+09           Max.   :330.00   Max.   :40.00

  Day.charge      Eve.mins        Eve.calls       Eve.charge      Night.mins      Night.calls
Min.   : 0.00   Min.   :  0.00   Min.   : 0.000   Min.   : 0.00   Min.   :  0.00   Min.   :0.000
1st Qu.:26.77   1st Qu.: 32.00   1st Qu.: 3.000   1st Qu.: 9.60   1st Qu.: 32.75   1st Qu.:2.000
Median :36.60   Median : 58.00   Median : 4.000   Median :17.40   Median : 58.00   Median :3.000
Mean   :40.67   Mean   : 78.65   Mean   : 5.081   Mean   :23.60   Mean   : 81.34   Mean   :3.459
3rd Qu.:53.70   3rd Qu.:120.00   3rd Qu.: 6.000   3rd Qu.:36.00   3rd Qu.:125.00   3rd Qu.:5.000
Max.   :99.00   Max.   :236.10   Max.   :35.000   Max.   :70.83   Max.   :252.00   Max.   :9.000

 Night.charge      Int.mins       Int.calls      Int.charge      Customer.ser   Churn    Occupation
Min.   : 0.000   Min.   : 0.00   Min.   :0.00   Min.   :  0.00   Min.   :0.000   N:55   Student : 8
1st Qu.: 3.275   1st Qu.: 0.00   1st Qu.:0.00   1st Qu.:  0.00   1st Qu.:1.000   Y:19   Business: 7
Median : 5.800   Median :10.00   Median :1.00   Median : 30.00   Median :2.000          CA      : 6
Mean   : 8.134   Mean   :15.74   Mean   :1.27   Mean   : 42.51   Mean   :2.311          Police  : 5
3rd Ou.:12.500   3rd Ou.:28.00   3rd Ou.:2.00   3rd Ou.: 67.50   3rd Ou.:3.000          Teacher : 5
```

C. Classification tree for all types of calls fit<- rpart(Churn~Customer.ser+Day.calls+Night.calls+ Eve.calls+Int.calls, method="anova", data=data) plot (fit, uniform=TRUE, main="Classification tree for data")

text (fit, use.n=TRUE, all=TRUE, cex=.7)

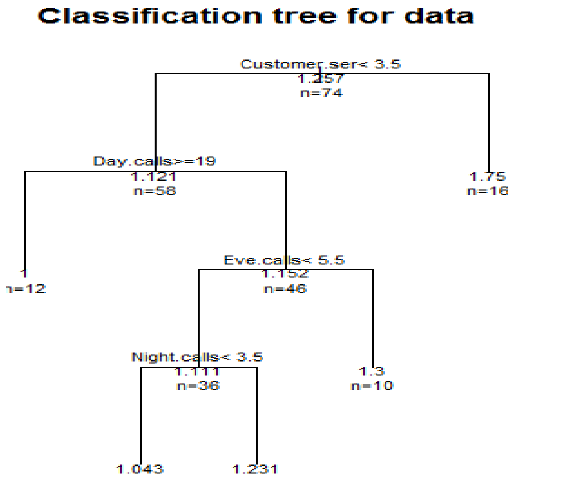


Figure 3: Represents the classification tree considered in churn data set

D. Using Plot function

Figure 4 represents the graph of age with respect to churn factor and area with respect to churn factors that has two values yes (Y) and no (N)
plot (Churn ~., data=data, type="c")

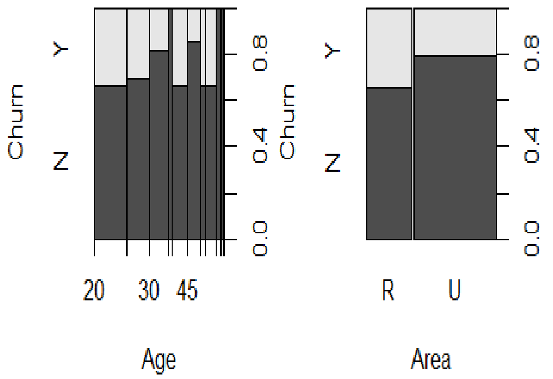


Figure 4: Age and Area with respect to churn factor

Figure 5 represents the graph of total number of calls with respect to churn factor using plot function.

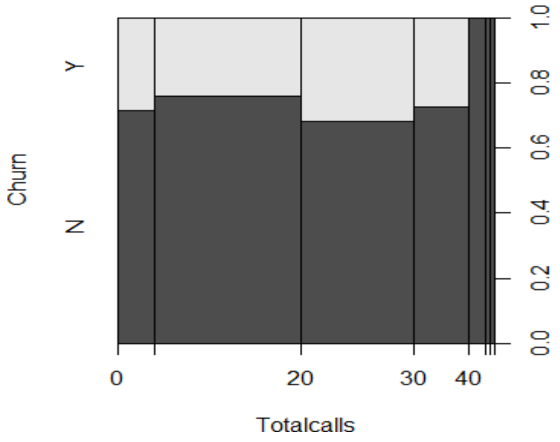


Figure 5: Total calls with respect to churn factor

Figure 6 Represents the set of all possible cost complexity Pruning's of a tree from a nested set. For the geometric Means of intervals of values of cp for which a pruning is Optimal, a cross validation has been done in initial Construction by rpart. The cp table in the fit contains the mean And standard deviation of the errors in the cross validation prediction against the each of the geometric means, and these are plotted by this function. A good choice of cp for pruning is often the leftmost value for which the mean lies below the horizontal Line.

fit<-rpart (Churn~Customer.ser+Day.calls+Night.calls+Eve.calls+Int.calls, method="anova", data=data) plotcp(fit)

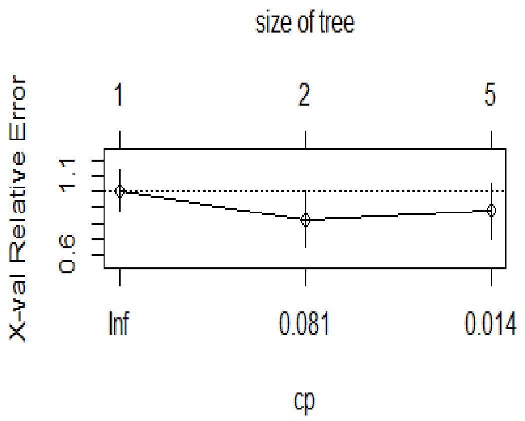
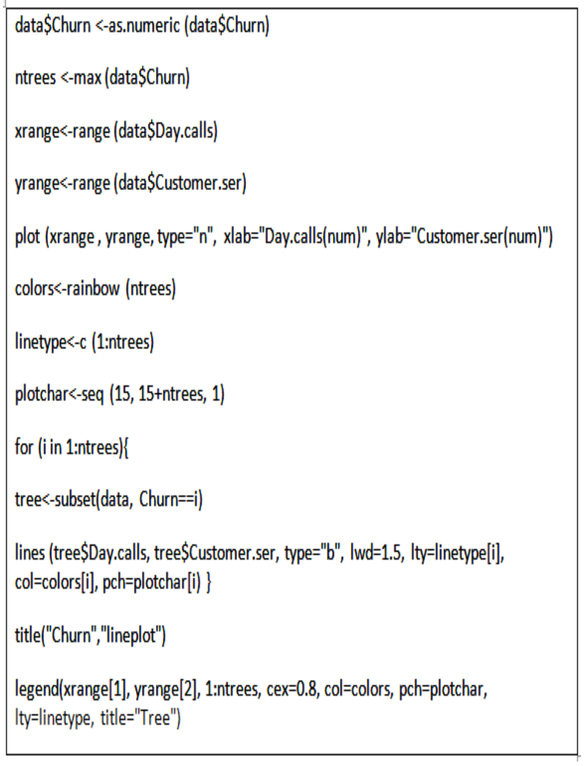


Figure 6: plotcp function

Figure 7 shows the line chart of Day calls and customer service calls using number as a range and considering the churn factor. The number of churn increased with the number of increase in customer service calls.

Line chart for Day calls and Customer Service calls



```
data$Churn <-as.numeric (data$Churn)

ntrees <-max (data$Churn)

xrange<-range (data$Day.calls)

yrange<-range (data$Customer.ser)

plot (xrange , yrange, type="n", xlab="Day.calls(num)", ylab="Customer.ser(num)")

colors<-rainbow (ntrees)

linetype<-c (1:ntrees)

plotchar<-seq (15, 15+ntrees, 1)

for (i in 1:ntrees){

tree<-subset(data, Churn==i)

lines (tree$Day.calls, tree$Customer.ser, type="b", lwd=1.5, lty=linetype[i],
col=colors[i], pch=plotchar[i) }

title("Churn","lineplot")

legend(xrange[1], yrange[2], 1:ntrees, cex=0.8, col=colors, pch=plotchar,
lty=linetype, title="Tree")
```

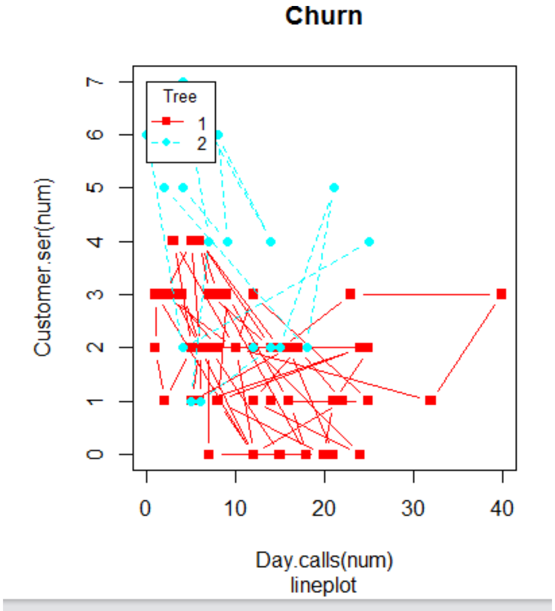


Figure 7: Line chart with respect to day calls and customer service calls

Figure 8 shows the line chart of age and night calls using number as range and considering factor. The number of night calls increases with decreased age group

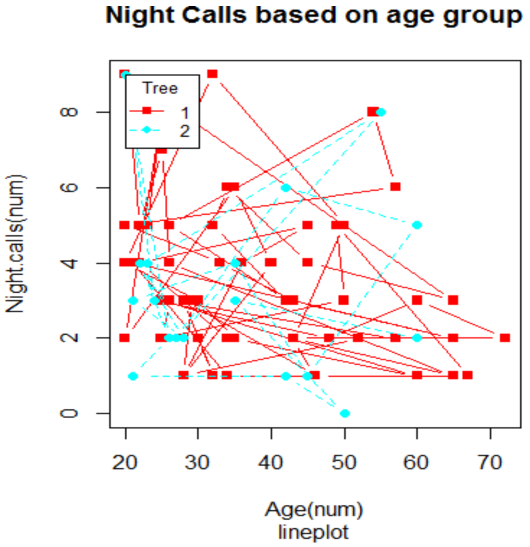


Figure 8: Line chart with respect to age and Night calls

Figure 9 represents the two curves on a single part in R. First curve (blue) corresponding to the Urban area and second curve (red) corresponding to Rural area

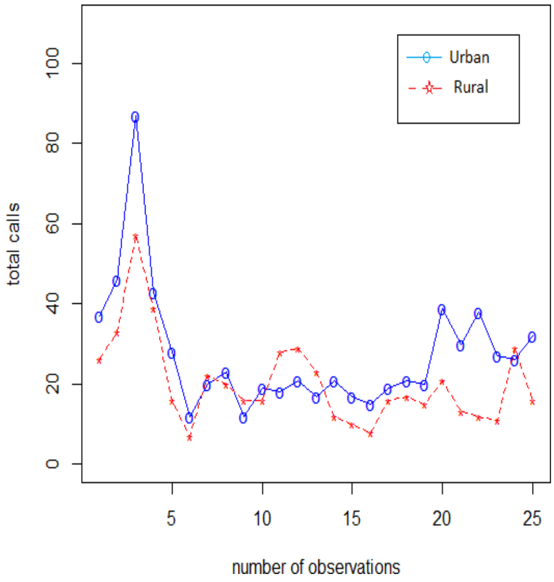


Figure 9: Line graph of the total number of calls with respect to Urban and Rural areas

The above line graph indicates that the total number of calls in urban areas is more than the rural areas due to lack of knowledge about the available services and offers providing by the telecom Industries.

rsq.rpart () plots the approximate R-square for the different splits.

It produces two plots. The first plots the r-square (apparent and apparent-from cross-validation) versus the number of splits. The second plots the Relative Error (cross-validation) from the cross- validation versus the number of splits.

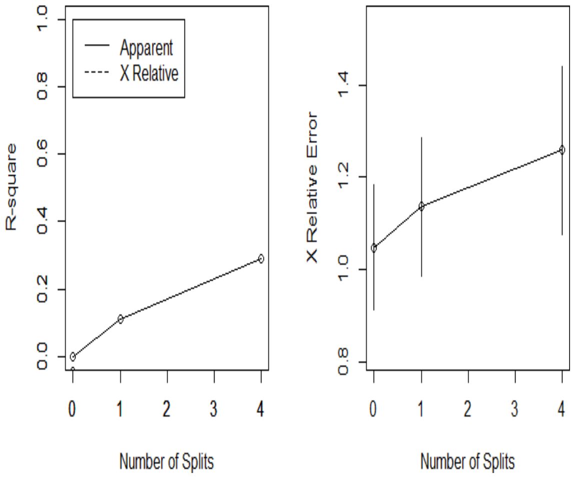


Figure 10: Approximate R-square for the different splits.

V. CONCLUSION

The current needs of telecom companies are a tool which can be used to understand customer behaviour and locate churners and to take possible actions to convert the churners to non-churners. The present paper focuses on developing a system which can be easily adopted by telecom Industry to discover the customer patterns, make predictions, Identify possible churners and to increase the net profit. The study predicts that the there is a huge deviations in the line graph of churners by measuring with respect to the customer service calls.

REFERENCES

[1] Manpreet Kaur, Dr. Prerna Mahajan, “Churn Prediction in Telecom Using R”, International Journal of Engineering and Technical Research (IJER), ISSN: 2321-0869, Volume-3, Issue-5, May 2015.

[2] Pushpa, G Shoba, “Social Network Classifier for Churn Prediction in Telecom Data”, 2013 International Conference on Advanced Computing and Communication Systems (ICACCS-2013), Dec. 19-21, Coimbatore, INDIA.

[3] “Applying Data Mining Techniques in Telecom Churn Prediction”, International Journal of Advanced Research in Computer Science and Software Engineering, N.Kamalraj Dr. A. Malathi, Department of Computer Technology PG and Research department of computer science, Dr. SNS Rajalakshmi College of Arts and Science Govt. Arts College.

[4] Dr. M. Balasubramanian, M. Selvarani, “Churn Prediction in mobile Telecom System using Data mining Techniques”, Department of Computer Science, Annamalai University, Chidambaram, International Journal of Scientific and Research Publication.

[5] Vladislav Lazarov, Marius Capota, “Churn Prediction”, Techsche Universitat Munchen.

[6] Chung fang Zhao, Yingliang Wu, HaijunGao “Study on knowledge Acquisition of the Telecom Customer” Consuming Behaviour Based on Data Mining”, 2008 IEEE.