Churn Prediction in Telecom Industry Using R

**Shwetha S Patil, Aniq Khan, Komal Saraya, Shwetha S N**

**Student, CSE, MSRIT, Bangalore, India**

**Abstract**

Telecommunication market is expanding loss of revenue due to increasing competition hence the loss of customers. They are trying to find the reasons of losing customers by measuring customer loyalty to regain the lost customers. The customers leaving the current company and moving to another telecom company are called churn.

The research paper is using data mining technique and R package to predict the results of churn customers on the benchmark Churn dataset available from (http://www.dataminingconsultant.com/data/churn.txt). The R tool has represented the large dataset churn in form of graphs which depicts the outcomes in various unique pattern visualizations. The Churn Factor is used in many functions to depict the various areas or scenarios where churners can be distinguished. The paper is considering churn factor in account to depict various patterns for churners. R is a powerful statistical programming tool which can represent the dataset graphically with respect to different parameters and it also uses different packages available.

Churns can be reduced by analyzing the past history of the potential customers systematically. In the past few years, the fast emerging requirements from both academia and industry has helped R programming language to emerge as one of the necessary tool for visualization, computational statistics and data science

**Keyword**- Churn, R Tool, Telecommunication.

**I. INTRODUCTION**

Numerous telecom companies are present all over the world. Telecommunication market is facing a severe loss of revenue due to increasing competition among them and loss of potential customers. Many companies are finding the reasons of losing customers by measuring customer loyalty to regain the lost customers. To keep up in the competition and to acquire as many customers, most operators invest a huge amount of revenue to expand their business in the beginning. Therefore, it has become important for the operators to earn back the amount they invested along with at least the minimum profit within a very short period of time.

* 1. **Churn Prediction**

Churn in the terms of telecommunication industry are the customers leaving the current company and moving to another telecom company. With the increasing number of churns, it becomes the operator‘s process to retain the profitable customers known as churn management. In telecommunication industry each company provides the customers with huge incentives to lure them to switch to their services, it is one of the reasons that customer churn is a big problem in the industry nowadays. To prevent this, the company should know the reasons for which the customer decides to move on to another telecom company. It is very difficult to keep customers intact for long duration as they move to the service that suits most of their needs.

* 1. **Types**

Telecom Churns can be classified in two main categories: Involuntary and Voluntary. Of the two, Involuntary are easier to identify. Involuntary churn are those customers whom the Telecom industry decides to remove as a subscriber. They are churned for fraud, non-payment and those who don‘t use the service. On the other hand, Voluntary churn are difficult to determine, here it is the decision of the customer to unsubscribe from the service provider. Voluntary churn can further be classified as incidental and deliberate churn. The former occurs without any prior planning by the churn but due to change in the financial condition, location, etc. Whereas, the latter happens for technological advancement, economics, quality factors and convenience reasons. Most operators are trying to deal with these type of churns mainly

* 1. **Managing Churns**

Churn management is very important for reducing churns as acquiring a new customer is more expensive than retaining the existing ones. Churn rate is the measurement for the number of customers moving out and in during a specific period of time. If the reason for churning is known, the providers can then improve their services to fulfill the needs of the customers. Churns can be reduced by analyzing the past history of the potential customers systematically. Large amount of information is maintained by telecom companies for each of their customers that keeps on changing rapidly due to competitive environment. This information includes the details about billing, calls and network data. The huge availability of information arises the scope of using Data mining techniques in the telecom database. The information available can be analyzed in different perspectives to provide various ways to the operators to predict and reduce churning. Only the relevant details are used in analysis which contribute to the study from the information given. Data mining techniques are used for discovering the interesting patterns within data. One of the most common data mining technique is Classification, its aim is to classify unknown cases based on the set of known examples into one of the possible classes. Here, in case of telecom churn, Classification helps learn to predict whether a customer will churn or not based on customer‘s data stored in database.

**II. BACKGROUND**

**2.1. Data Mining Techniques**

The process of reducing, analyzing the patterns, predicting the hidden and useful required information from large Database is known as Data Mining. Association rule mining, clustering, classification and regression forms the four techniques used by data mining.

In Data mining new rules and patterns can be discovered by the system known as discovery oriented and system can also check the user‘s hypothesis called verification oriented. It helps in taking knowledge-driven decisions and for predicting the future trends of the business.

**2.2. Tool Used: A Revolution Analytics Tool - R**

In the past few years, the fast emerging requirements from both academia and industry has helped R programming language to emerge as one of the necessary tool for visualization, computational statistics and data science. R is most popular in field of data science and important in Finance and analytics- driven companies.

R virtually consists all the possible statistical models, data manipulation and charts that could ever be required by a modern day scientist. One can easily use the best reviewed methods from leading researchers in field of Data Science without any cost. It provides a large collection of graphical and statistical techniques, consisting of modelling (linear and non-linear), statistical tests, time-series, classification, clustering, etc.

R helps in representing complex data as beautiful and unique data visualizations. Evaluation of result in R is very much easier as we do not have to remember any clicks or steps, it is simply a programming language designed specifically for data analysis that also has the capability to use mix and match models for best results.

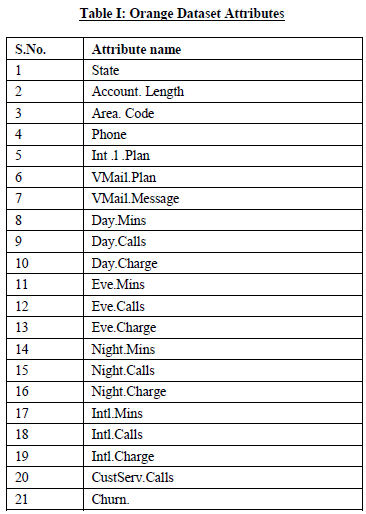
As R is supported by a large community worldwide, solution to the errors and code is available freely. Its source code is written in C, Fortran and R. R is easily extensible through functions and extensions, and the

R community is noted for its active contributions in terms of packages. R is an open source and can be extended easily as individuals using it can contribute in its growth. Dynamic and static graphics are available through additional packages. R can easily deal with complex and large datasets.

The libraries and packages of R that are being used in this paper are: rpart, class

**2.3 Data Set Used**

The attributes in our data are taken from Orange Database.



**III. ALGORITHM AND LIBRARIES USED**

**3.1 Extracting Classification Rules from Trees**

1. IF-THEN rules are used.

2. From root to leaf one rule is created for each path.

3. Each attribute-value pair along a path forms a conjunction.

4. The leaf node holds the class prediction.

5. Rules are easier for humans to understand.

**3.3. Using rpart package**

rpart (formula, data, method)

f<-rpart(Churn.~CustServ.Calls+Eve.Charge+Intl.Charge+Night.Charge+Day.Charge, method="class", data=churn)

Package rpart is used in plotting the graphs. The functions within rpart that are used are as follows:

**3.4 Using plotcp function**

Applied on the set of possible cost- complexity pruning of a tree from a nested set. A cross- validation is already performed by rpart on the geometric means of the Interval values of cp where pruning is optimal. The mean and standard deviation of errors in cross- validated prediction against each of the geometric means is stored in cptable in ‗f‘ 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.

**3.5. Using Plot function**

plot(Churn. ~., data = churn, type = "c")

lines(Churn.~ Day.Charge,type="l")

In plot function, x and y axis are mentioned along with the data source and the type of graph that is, curve, line etc.

**IV. EXPERIMENTAL RESULTS AND OBSERVATION**

**4.1 Reading Data Set Churn from CSV file**

churn<-read.csv("C:\\Users\\Documents\\R\\win-library\\3.1\\RWeka\\R\\churn.csv", header=T)

**4.2. Names of all the attributes**

> names(churn)

[1] "State" "Account.Length" "Area.Code" "Phone"

[5] "Int.l.Plan" "VMail.Plan" "VMail.Message" "Day.Mins"

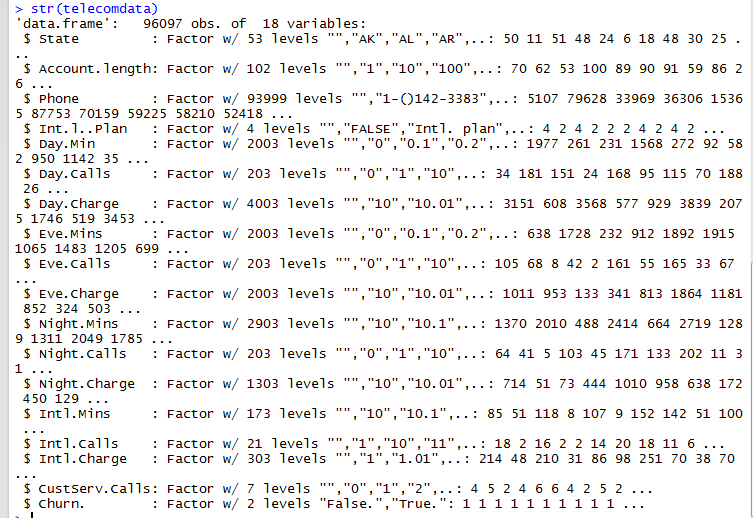
[9] "Day.Calls" "Day.Charge" "Eve.Mins" "Eve.Calls"

[13] "Eve.Charge" "Night.Mins" "Night.Calls" "Night.Charge"

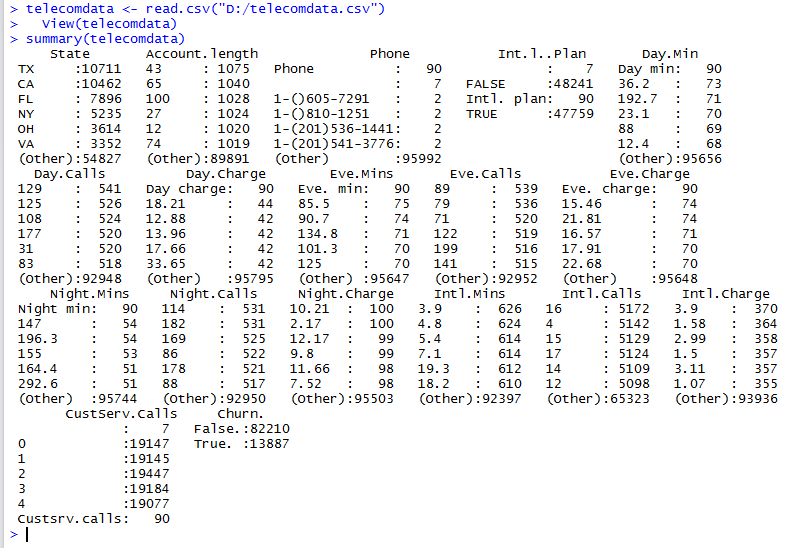
[17] "Intl.Mins" "Intl.Calls" "Intl.Charge" "CustServ.Calls"

[21] "Churn."

**4.3. Description of complete data Set**



**4.4. Summary of Data set**

****

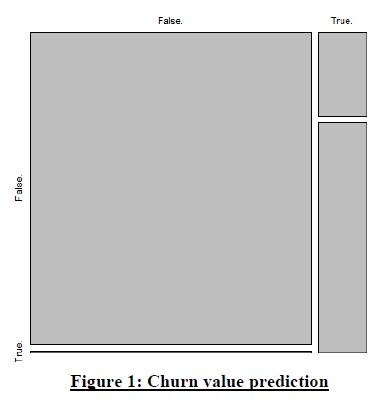
**4.3. Classification Tree for all the Calls (using rpart )**

library(rpart)

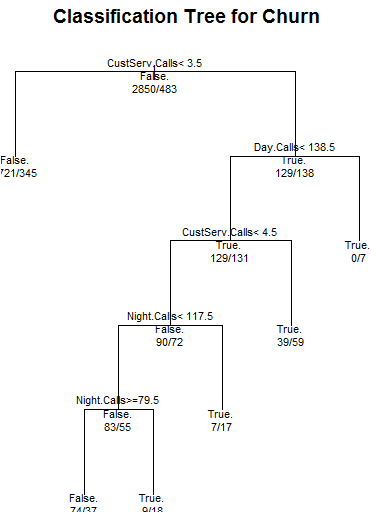
f<-rpart(Churn.~CustServ.Calls+Eve.Calls+Intl.Calls+Night.Calls+Day.Calls,method="class", data=churn)

plot(f, uniform=TRUE,main="Classification Tree for Churn")

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



**Fig. 1** depicts the churn values from table formed by predicting the values of J48 decision tree on churn parameter.



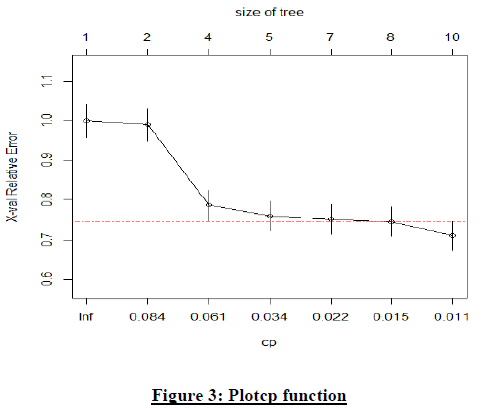
**Fig. 2** represents the classification tree for all the Calls considered in churn Dataset. The decision is made on basis of call number and the churn factor having values true and false

**4.4. Using rpart**

library(rpart)

f<-rpart(Churn.~CustServ.Calls+Eve.Charge+Intl.Charge+Night.Charge+Day.Charge, method="class", data=churn)

plotcp(f,lty=4,col="red")



**Fig. 3** represents Applied on the set of possible cost- complexity pruning of a tree from a nested set. A cross- validation is already performed by rpart on the geometric means of the Interval values of cp where pruning is optimal. The mean and standard deviation of errors in cross- validated prediction against each of the geometric means is stored in cptable in ‗f‘ are plotted by this function.

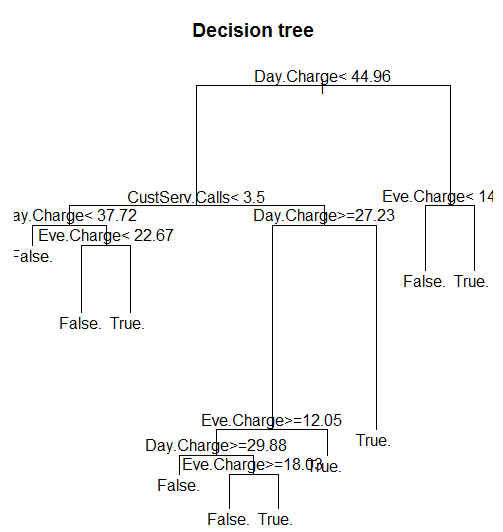
**4.5. Dicision tree**

library(rpart)

tree <- rpart(Churn. ~CustServ.Calls + Eve.Charge + Intl.Charge + Night.Charge + Day.Charge ,data = churn1 ,method = "class")

plot(tree ,main="Decision tree")

text(tree,pretty = 0)



**Fig 4: Decision Tree**

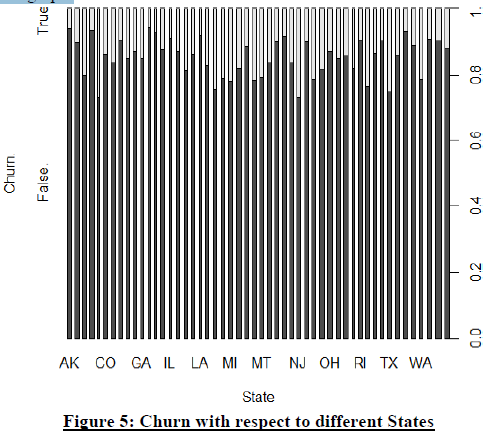
**Fig. 4** depicts the churn values from table formed by predicting the values of J48 decision tree on churn parameter

**4.6. Using Plot Function**

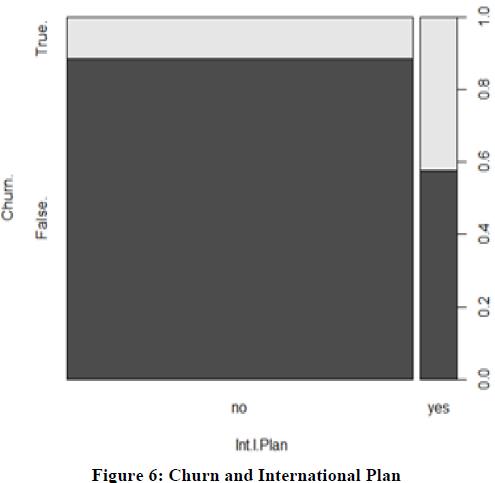
plot(Churn. ~., data = churn, type = "c")

lines(Churn.~ Day.Charge,type="l")

.



**Fig. 5**  represents the graph of churn factor with respect to all the states. The number of churns can easily be observed state wise in the graph.

****

**Fig. 6** represents the graph of churn factor with respect to International plan

**4.7 Logistic regression using logmodel**

logmodel <- glm(Churn. ~ CustServ.Calls + Eve.Charge + Intl.Charge + Night.Charge + Day.Charge ,data = churn1,family=binomial())

summary(logmodel)

**V. CONCLUSION**

The proposed research has used data mining technique and R package to predict the results of churn customers on the benchmark Churn dataset available at http://www.sgi.com/tech/mlc/db/ and http://www.dataminingconsultant.com/data/churn.txt. It has evaluated, the number of churns using the classification technique Decision Tree. The R tool has represented the large dataset churn in form of graphs which depicts the outcomes vividly and in a unique pattern visualization manner. The Churn Factor is used in many functions to depict the various areas or scenarios when the churn rate is high. The study predicts that there is a huge deviation in graph of churners when customer service calls are measured. The graphs are made taking churn factors as the deciding parameters. Graphs represent the different ways of observing the number of churners from the dataset. Once the root area is recognized the steps can be taken by Telecom Company to improve their services and retain their old customers from churning

**REFERENCES**

[1] ―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.

[2] Anshul Goyal and Rajni Mehta, Assistant Professor,CSE. ―Performance Comparison of Naïve Bayes and J48 Classification Algorithms.‖

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

[4] Vladislav Lazarov, Technische Universität München and Marius Capota, Technische Universität München, ―Churn Prediction‖,

[5] Prof S. Chandrasekhar, ―Predicting The Churn In Telecom Industry‖‘.

[6] Dr. Rajni Jain, ―Introduction to Data Mining Techniques‖.

[7] Marcin Owczarczuk, ―Churn models for prepaid customers in the cellular telecommunication industry using large data marts‖, Expert Systems with Applications, 37 (2010) 4710–4712.

[8] Gary M. Weiss, ―Data Mining in the Telecommunications Industry‖, Data Mining and Knowledge Discovery Handbook, 2005 - Springer

[9] ―Data Mining And Business Analytics with R‖, Johannes Ledolter, Dept. of Management Studies, University of IOWA.