PERFORMANANCE ENHANCEMENT OF CUSTOMER CHURN PREDICTION IN TELECOM SECTOR USING DECISION TREE TECHNIQUES

Priyanka A. Shrikhande
Department of Computer Science
Sagar Institute Of Research & Technology-Excellence
Bhopal,India

Prof. Anjana Verma
Department of Computer Science
Sagar Institute Of Research & Technology-Excellence
Bhopal,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 decision tree techniques such as CART and Random Forest for predicting customer churn through which we can build the classification models and also compare the performance of these models with logistic regression model.**

Keywords—Churn prediction, data mining, telecom data,Customer retention, classification, decision tree, CART, Random Forest.

I. INTRODUCTION

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.[6]

The Churn Analysis [4] 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[7], 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 effect 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

According to [1],With the development and popularization of Internet technology, e-commerce platform has provided satisfying products for customers and cultivated customer loyalty. Nevertheless, the loss of user is still a popular issue in business field and academic field. Based on logistic regression model, this paper established an e-commerce user churn prediction model through preliminary research on e-commerce customer churn behavior. By using the factor analysis method, the user's online duration, number of logins, attentions, and other user behavior factors were analyzed which concludes the factor affecting the loss of users. Finally, the empirical study proved that the proposed EBURM model can predict user churn behavior in a high confidence level. Through the construction of the EBURM model to predict ecommerce user churn

behavior, it helps e-commerce platform to formulate operational strategy more precisely, provide users with personalized recommendations, increase user activity, retain users, and improve the economic effects of e-commerce platform.

In [2], Since the beginning of data mining the discovery of knowledge from the Databases has been carried out to solve various problems and has helped the business come up with practical solutions. Large companies are behind improving revenue due to the increase loss in customers. The process where one customer leaves one company and joins another is called as churn. This paper will be discussing how to predict the customers that might churn, R package is being used to do the prediction. R package helps represent large dataset churn in the form of graphs which will help to depict the outcome in the form of various data visualizations. Churn is a very important area inwhich the telecom domain can make or lose their customers and hence the business/industry spends a lot of time doing predictions, which in turn helps to make the necessary business conclusions. Churn can be avoided by studying the past history of the customers. Logistic Regression is been used to make necessary analysis. To proceed with logistic regression we must first eliminate the outliers that are present, this has be achieved by cleaning the data (for redundancy, false data etc) and the resultant has been populated into a prediction excel using which the analysis has been performed.

According to [3] , A big problem that encounters businesses, especially telecommunications business is 'customer churn'; this occurs when a customer decides to leave a company's landline business for another cable competitor. Therefore, our aim beyond this study to build a model that will predict churn customer through defining the customer's precise behaviors and attributes. We will use data mining techniques such as clustering, classification and association rule. The accuracy and preciseness of the technique used is so essential to the success of any retention attempting. After all, if the company is not aware of a customer who is about to leave their business; no proper action can be taken by that company towards that customer.

In the churn analysis paper [13], 4 different core functions have been used in the Support Vector Machines model and performances have been compared by using a data set consisting of 3333 customer records with 21 variables provided by a telecommunications company. And among these models, the one with the polynomial core function has been reported to have the best result by 88.56%.

In paper [14] theydiscusses a homophily-based customer churn analysis implementation, described as the tendency of individuals to demonstrate similar behavior to those in their social network environment, as well as investigating the details of phone call records of individuals. They have used a data set of 6 months' call records of 1 million customers with 111-variable including data such as which customer had a conversation with whom, how many times, and for how long, which is used to depict their social network provided by a GSM operator. They have stated that the test results of these combined factors were more successful than of the individual test cases.

III. PROBLEM DEFINITION

From the problems obligatory through market saturation and value implications, there has been associate identification of a desire for a computer based mostly churn prediction methodology that's capable of accurately distinctive a loss of client ahead, so proactive retention ways is deployed during a bid to retain the client. The churn prediction should be correct as a result of retention ways is pricey. A limitation of current analysis is that alternative studies have focused virtually solely on churn capture, neglecting the problem of misclassification of non-churn as churn. Retention campaigns usually embrace creating service based mostly offers to customers during a bid to retain them.

IV. PROPOSED WORK

In the proposed system R [8] programming will be used to build the model for churn prediction. It is widely used among statisticians and data miners for developing statistical software and data analysis. R is freely available and a powerful statistical analysis tool which has not yet been explored for building models for churn prediction[3].

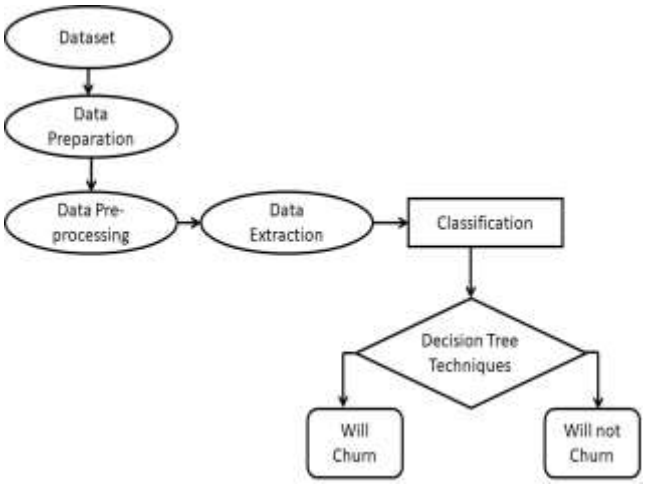

Figure1. Churn Prediction Framework

In this paper, we proposed different decision tree algorithms to analyze customer churn analysis. Through which we can multiple different models are employed to accurately predict those churn customers in the data set. These models are CART and Random Forest.

Our Steps or Algorithm Steps will follow:

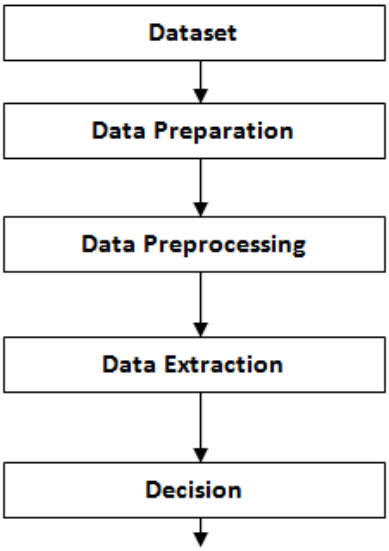

Figure 2. Analysis Steps

1. Dataset:- A telecom dataset is taken for predicting churn which to identify trends in customer churn at a telecom company and the data which we taken is in .csv format. The data given to us contains 3333 observations and 21 variables extracted from a datasets.

2. Data Preparation: Since the dataset acquired cannot be applied directly to the churn prediction models, so we can naming each attributes.

3. Data Preprocessing: Data preprocessing is the most important phase in prediction models as the data consists of ambiguities, errors, redundancy and transformation which needs to be cleaned beforehand.

4. Data Extraction: The attributes are identified for classifying process.

5. Decision: Based on data extraction and classification models we can take a decision whether the employee is churner or not.

V. EXPERIMENTAL & RESULT ANALYSIS

All the experiments were performed using an i5-2410M CPU @ 2.30 GHz processor and 4 GB of RAM running Windows. After that we can install r base core on windows and Rstudio and then to identify trends in customer churn at a telecom company. The data given to us contains 3,333 observations and 21 variables extracted from a data warehouse. These variables are shown in figure 2.

S.No.	Attribute name
1	State
2	Account. Length
3	Area. Code
4	Phone
5	Int .l .Plan
6	VMail.Plan
7	VMail.Message
8	Day.Mins
9	Day.Calls
10	Day.Charge
11	Eve.Mins
12	Eve.Calls
13	Eve.Charge
14	Night.Mins
15	Night.Calls
16	Night.Charge
17	Intl.Mins
18	Intl.Calls
19	Intl.Charge
20	CustServ.Calls
21	Churn.

Figure 3. Telecom dataset attribute

Than we build our classification model based on classification and regression tree (CART) algorithm.So the model first perform feature selection method in which its select the best attributes which is used to create a

decision tree. Figure 3 show the classification tree result.

```
> treemodel <- tree(Churn, ~.-State-Phone, data = telecom)
> summary(treemodel)
> summary(treemodel)

Classification tree:
tree(formula = Churn. ~ . - State - Phone, data = telecom)
Variables actually used in tree construction:
[1] "Day.Mins"      "CustServ.Calls" "Int.l.Plan"     "Eve.Mins"
[5] "VMail.Plan"    "Intl.Calls"     "Intl.Mins"
Number of terminal nodes:  12
Residual mean deviance:  0.3772 = 1253 / 3321
Misclassification error rate: 0.05911 = 197 / 3333
>
```

Figure 4. Classification tree

In figure 3 we get the classification tree result, in which the model choose 7 variables which are actually used in tree construction. These attribute selection is depend on values contains by attributes because these values may be continuous or categorical so the attribute selection method calculate the gini index of these attributes and select the best attributes through which we can construct the decision tree. Then we use the plot function and plot the classification decision tree which is shown in figure 4.

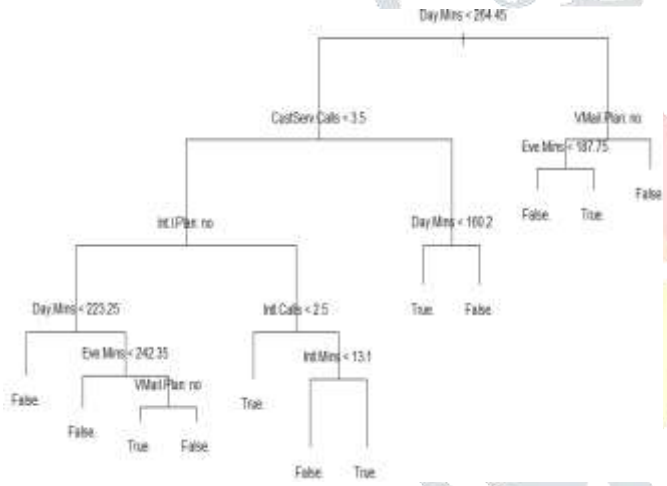

Figure 5. Decision tree

Based on the decision tree the model predicts the outcomes in two category false (non churn) or true (churn). So we can compare the predicted outcomes to the actual outcomes and the outcomes are shown in figure 5.

```
Console ~/
          Actual
Predicted False. True.
   False.   2811   158
   True.      39   325
>
> Accuracy <- print((cm[2,2]+cm[1,1])/sum(cm) * 100)
[1] 94.08941
>
> Sensitivity<-print(cm[2,2]/(cm[2,2]+cm[1,2])*100)
[1] 67.28778
>
> Specificity<-print(cm[1,1]/(cm[1,1]+cm[2,1])*100)
[1] 98.63158
>
```

Figure 6. CART outcomes

In the output we can seen that there are predicted values comes along with actual value, means in that we predicted right that 2811 customers cannot churn, and we predicted 325 peoples are churn. And based on the confusion matrix we can calculate the performance measure of CART model.

After computing performance of CART model we explore an another decision tree based algorithm which is random forest. The CART model predict the outcomes based on the single tree whereas random forest takes multiple recursive tree for predicting the outcomes. We can train the random forest model by applying telecom dataset to these model by providing 75% data for training and 25% for testing , figure 6 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")
```

Figure 7.Random forest classifier

After the model gets trained we can test the model on testing dataset and the outcomes which we get are shown in figure 7.

```
Console ~/
          Reference
Prediction False. True.
    False.   84.8   4.0
    True.     1.0  10.3

 Accuracy (average) : 0.951

>
> pred<-predict(rfmodel, newdata=churnTest)
> confusionMatrix(pred, churnTest$Churn.)
Confusion Matrix and Statistics

          Reference
Prediction False. True.
    False.   700    36
    True.     12    84

               Accuracy : 0.9423
                 95% CI : (0.9242, 0.9572)
    No Information Rate : 0.8558
    P-Value [Acc > NIR] : 1.849e-15

                  Kappa : 0.7451
 Mcnemar's Test P-Value : 0.0009009

            Sensitivity : 0.9831
            Specificity : 0.7000
         Pos Pred Value : 0.9511
         Neg Pred Value : 0.8750
             Prevalence : 0.8558
         Detection Rate : 0.8413
   Detection Prevalence : 0.8846
      Balanced Accuracy : 0.8416

       'Positive' Class : False.

>
```

Figure 8. Random Forest outcomes

CROSS-VALIDATION

Confusion Matrix

From the confusion matrix there are exactly 4 possible outcomes from a binomial classifier model. The total number of positive instances in the matrix is $T = FP + TP$ and the total number of negative instances is $F = TN + FN$. The most common evaluation metrics are overall accuracy, true positive rate and false positive rate.

		Predicted classes	
		Class=Yes/+/ Churn	Class=No/-/No-churn
Actual classes	Class=Yes/+/ Churn	TP (true positive)	FN (false negative)
	Class=No/-/ No-churn	FP (false positive)	TN (true negative)

Accuracy= TP+TN/TOTAL

Sensitivity = TP/(TP+FN)

Specificity= TN/(TN+FP)

We can compute the classifier performance measure which we get from confusion matrix and the measures of CART and Random forest and compare the performance of these models with Logistic regression model [1] and the comparison result are shown in table-1.

Techniques	Accuracy	Specificity	Sensitivity
LOGISTIC REGRESSION	93.60	96.03	95.22
CART	94.08	98.63	67.29
RANDOM FOREST	95.10	70.00	98.31

Table 1. Classification measures

When comparing these result into the above table it is clear that we can the predicting churn is more accurate in random forest because it gives 98.31% sensitivity. The true positive rate known as the sensitivity or probability of detection measures the proportion of positives that are correctly identified as churn.

Figure9. Accuracy Comparison

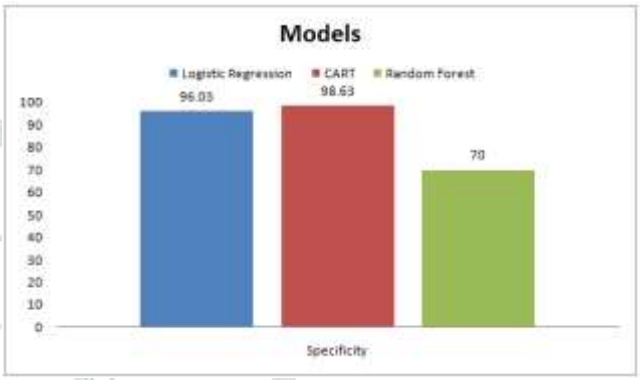

Figure 10. Specificity Comparison

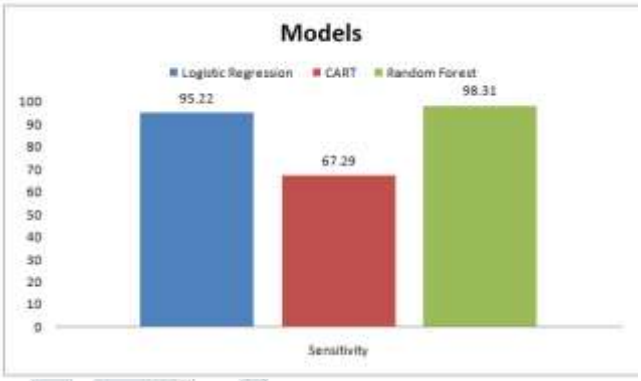

Figure 11. Sensitivity Comparison

VI. CONCLUSION

In order to retain existing customers, Telecom providers need to know the reasons of churn, which can be realized through the knowledge extracted from Telecom data.In this paper, we builds a classifiers based on decision tree techniques such as CART and Random Forest and compare the performance with Logistic Regression model. Based on the result we can say that random forest perform better in customer churn prediction because it has better sensitivity as compared to CART and Logistic Regression models. Future work will be applying rules based techniques on telecom datasets and compare the results with some of the most commonly used techniques in churn prediction as they are very suitable tools for data mining applications.

REFERENCES

[01] QiuYanfang , Li Chen , "Research on E-commerce User Churn Prediction Based on Logistic Regression " in IEEE 2017.

[02] Helen Treasa Sebastian* and Rupali Wagh in "Churn Analysis in Telecommunication using Logistic Regression " in ORIENTAL JOURNAL OF COMPUTER SCIENCE & TECHNOLOGY, 2017.

[03] Ibrahim M.M.Mitkees , Asist. Prof. Sherif M Badr , Dr. Ahmed Ibrahim BahgatElSeddawy , in "Customer Churn Prediction Model using Data Mining techniques " in IEEE 2017.

[04] Rahul J. Jadhav, Usharani T. Pawar, "Churn Prediction in Telecommunication Using Data Mining Technology", in *(IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 2, No.2, February 2011*

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

[06] N.Kamalraj, A.Malathi' " A Survey on Churn Prediction Techniques in Communication Sector" in *International Journal of Computer Applications (0975 – 8887) Volume 64– No.5, February 2013*

[07] Kiran Dahiya,KanikaTalwar, "Customer Churn Prediction in Telecommunication Industries using Data Mining Techniques- A Review" in International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 4, 2015

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

[09] Data Mining in the Telecommunications Industry‖, Gary M. Weiss, Fordham University, USA.

[10] Argüden Y., Erşahin B. 2008. VeriMadenciliği: VeridenBilgiye, MasraftanDeğere. ARGE Danışmanlık, ISBN: 978-975- 93641-9-9 1. Basım.

[11] Kotler, P., Keller, K. L. 2009. Marketing Management. Pearson Prentice Hall.

[12] Seker, S.E. 2016. MüşteriKayıpAnalizi (Customer Churn Analysis). YBS Ansiklopedi, 3 (1): 26-29.

[13] Brandusoiu, I., Toderean, G. 2013. Churn Prediction In The Telecommunications Sector Using Support Vector Machines. Annals Of The Oradea Un., Fascicle Manag. and Tech. Eng., 1: 19-22.

[14] Backiel, A., Verbinnen, Y., Baesens, B., Claeskens, G. 2015. Combining Local and Social Network Classifiers to Improve Churn Prediction. International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 651-658.