PREDICTING CUSTOMER CHURN IN TELECOM SECTOR USING CLASSIFICATION AND REGRESSION TREE (CART)

Neelima Pandey
M. Tech Scholar,
OIST, Bhopal (M P)

Bhupendra Panchal
Asst. Prof. CSE,
OIST, Bhopal (M P)

Dr. Farha Haneef
HOD, CSE
OIST, Bhopal (M P)

***Abstract*— Predicting customer churn in telecommunication industries becomes a most important topic for research in recent years. Because its helps in detects which type of customer are likely to change their subscriptions to a particular service. Analysis of data which is extracted from telecom companies can helps to find the reasons of customer churn and also uses the information to retain the customers. So predicting churn is very important for telecom companies to retain their customers. This paper we build the classification model using Decision tree and evaluates the performance measures and compare it performance with logistic regression model.**

***Keywords*—** Classification, Churn prediction, telecom data, Logistic Regression model, Customer retention, CART Algorithm**.**

I.INTRODUCTION

Data mining strategies lie at the intersection of computing, statistics and machine learning info systems. Data processing techniques helps in building the prediction models to get future developments and actions permitting the organizations to require good selections derived from the data from knowledge [4].

Churn prediction is associate application of client performance in data processing. Churn could be a key issue sweet-faced through associate enterprise associated denoted the value of extending a replacement client is almost 5 times more than the value of maintaining an recent client. as a result of the fight of the enterprise market is declined through churn, churn prediction is dispensed through data processing to boost the client maintenance. corporations establish the customers Organization aren't passionless to maneuver close to a contender through churn prediction. After that, appropriate advertising operations square measure wont to preserve and hold the shoppers [6]. Churn prediction permits corporations to boost the potency of client retention operations and to reduce the prices joined with churn.

Churn prediction method on medium business is a vital analysis space for the popularity of the faults. client churn is characterized because the loss of shoppers as they leave to their competitors. it's key issue in medium industries as taking new customer’s [5]. In telecommunication business, the churn is additionally known as client attrition or subscriber churning. it's termed because the development of loss of a client. The method of movement from one supplier to a different is occurring due to the higher rates or services or different blessings that the contender company provides whereas language up.

Within the business setting, churn indicates client migration and loss of import. Churn rate is calculated because the share of shoppers Organization finish reference to the organization or with customers receiving their services. In new associations, there square measure giant demand that predicts the shoppers to take care of them promptly by reducing the prices and risks. It additionally will increase the potency and fight. they're employed in market advanced analytics tools and applications designed to spot the massive quantity of information within the teams. It additionally creates prediction derived from the knowledge earned by examining and exploring the information.

II. LITERATURE REVIEW

According (Qiu Yanfang, et. al., 2017) [1] , from the beginning of the data mining which is used to discover new knowledge's 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. This paper 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 [8]. In this logistic regression model are used 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 and get the various parameters like true positive rate, false positive rate and accuracy.

According to (Chuanqi Wang, et. al, 2017) [2] , telecommunication shopper churn prediction is additionally a cost sensitive classification. Most of studies regard it as a general classification use ancient ways in which that the 2 types of misclassification cost are equal. And, in facet of price sensitive classification, there are some researches targeted. They propose the partition cost-sensitive CART model throughout this paper. The experiment supported the required data, showed the maneuver not solely obtains an honest classification performance, however additionally reduces the full misclassification cost effectively.

According to (Duyen DO, et.al., 2017)[3], Telecom Organization have heaps of shoppers and maintain a good client relationship, will get substantial benefits. Their studies have shown that, inside the telecommunication business, the worth of effort new shopper is 5 to 6 times over mindful existing customers. Moreover, recent customers can generate higher profits. For this reason, they have studied and resolved consumer churn prediction.

III PROBLEM IDENTIFICATION

The problem is follows as:
Telecommunication corporations are presently not capable to predict corporations initiated churn.
How will telecommunication corporations uses data processing techniques and model selection techniques for churn prediction?
Client retention focuses on the subject of client churn, whereby churn pronounce earnings of shoppers, and management of churn designates efforts a business makes to spot and management the client churn drawback.

IV PROPOSED WORK

We proposed a machine learning classification techniques based on CART algorithm through which we can build the churn classification model for prediction. In this we uses R programming which is a open source language used for machine learning for building a models.

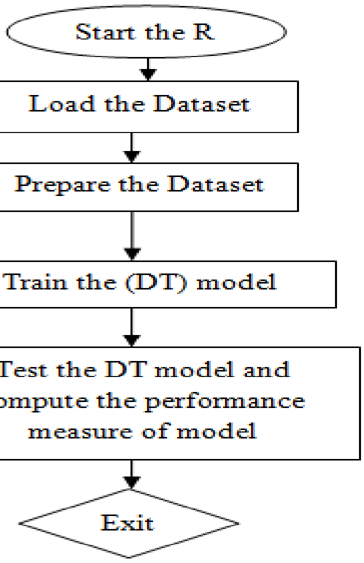


Figure 1.Flow Chart

1. Dataset:- A telecom dataset is taken for predicting churn which we taken is in .csv format. The dataset

consists of 20,000 observation (lines, rows) over 12 variables (fields, columns) reading features of customers of a mobile phone provider.

2. Data Preparation: we can naming each attributes. these attributes are similar as column names in which each attributes is a collection of similar type of record values.

3. Data Preprocessing: In these we can change the data type according to our need such as attributes consist value like true or false value so we can transform into 0 or 1 (numeric) data for our algorithm needs.

4. Data Extraction: we have worked with numerical and categorical values and from the 21 attributes the feature selection process select 7 attributes which is best for tree construction.

5. Decision: Based on these 7 attributes the tree model generates a decision tree. Now we have predicted value comes from decision tree model and the actual values so we can compare the predicted value with the actual values and create a confusion matrix and from these matrix values we can calculate the various performance measures like accuracy, specificity, and sensitivity.

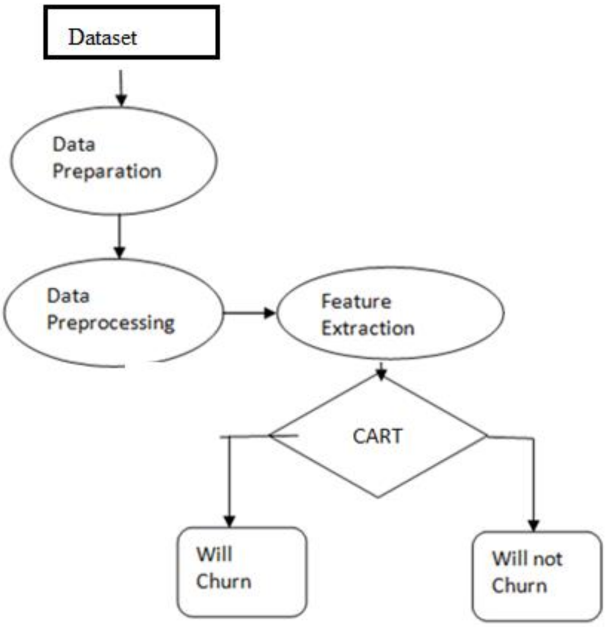


Figure 2. Churn Prediction Framework

Algorithm

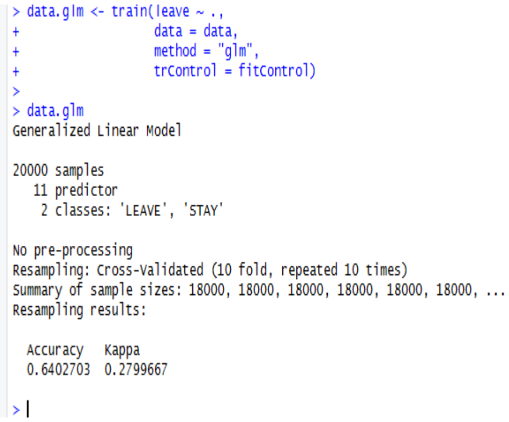
```
Input:
Data partition, D, which is a set of training tuples and their related class labels;
Attribute_list;
Attribute_selection_method, to determine the splitting criterion that "best" partitions.

Output: A decision tree.

Step 1- CREATING A ROOT NODE
1. Create a root node N
2. If tuples in D are all of the similar class, C then
3.      Return N as a leaf node label with the class C;
4. If attribute list is empty then
5.      Return N as a leaf node label with the majority class in D

Step 2- ATTRIBUTE SELECTION
6. Apply attribute_selection_method(D, attribute_list) to discover the "best " splitting _criterion attribute;
7. Label node N with splitting _criterion;
8. Update the attribute_list

Step 3- SPLIT THE TREE
9. for each outcome j of splitting_criterion
    //partition the tuples and produce subtrees for each partition
10.Based on splitting_criterion attribute
                Split the tree into two part
12.  attach a leaf labeled with the majority class D in node N:
13.      else attach the node returned by Generate_decision_tree(Dj:attribute_list)to node N:
      end for
14.  return N,
```

V EXPERIMENTAL & RESULT ANALYSIS

The dataset consists of 20,000 observation (lines, rows) over 12 variables (fields, columns) reading features of customers of a mobile phone provider, including the class variable LEAVE represent whether e customer decided to quit the service or not.

Logistic Regression Model

Let’s use the logistic regression for better understanding of which variables are important in the data set. We will use the bootstrapping method to train the model this time. We can preprocess the data for transform dataset for logistic regression model and then we are using glmnet package. After the model gets trained and we can compute the performance of the model and we gets the performance measure are shown in figure 3.

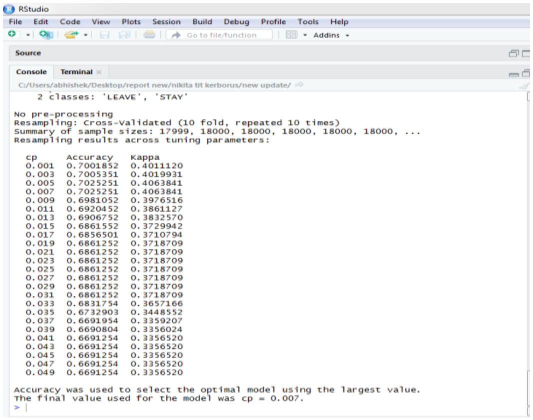


```
> data.glm <- train(leave ~ .,
+                   data = data,
+                   method = "glm",
+                   trControl = fitControl)
>
> data.glm
Generalized Linear Model

20000 samples
   11 predictor
    2 classes: 'LEAVE', 'STAY'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 10 times)
Summary of sample sizes: 18000, 18000, 18000, 18000, 18000, 18000, ...
Resampling results:

  Accuracy   Kappa
  0.6402703  0.2799667

> |
```

Figure 3. Performance measure of Logistic Regression

Decision Tree Model

Now we will take a decision tree model. We will use repeated cross validation to train the model with 10 folds and 10 repeats. This means that our model will essentially be training on 100 random subsections of our data. After the model gets trained and we can compute the performance of the model and we gets the performance measure are shown in figure 4.


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Source
Console
Terminal
C:/Users/abhishek/Desktop/report new/nikita tit kerborus/new update/

```
    2 classes: 'LEAVE', 'STAY'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 10 times)
Summary of sample sizes: 17999, 18000, 18000, 18000, 18000, 18000, ...
Resampling results across tuning parameters:

   cp     Accuracy   Kappa
   0.001  0.7001852  0.4011120
   0.003  0.7005351  0.4019931
   0.005  0.7025251  0.4063841
   0.007  0.7025251  0.4063841
   0.009  0.6981052  0.3976516
   0.011  0.6920452  0.3861127
   0.013  0.6906752  0.3832570
   0.015  0.6861552  0.3729942
   0.017  0.6856501  0.3710794
   0.019  0.6861252  0.3718709
   0.021  0.6861252  0.3718709
   0.023  0.6861252  0.3718709
   0.025  0.6861252  0.3718709
   0.027  0.6861252  0.3718709
   0.029  0.6861252  0.3718709
   0.031  0.6861252  0.3718709
   0.033  0.6831754  0.3657166
   0.035  0.6732903  0.3448552
   0.037  0.6691954  0.3359207
   0.039  0.6690804  0.3356024
   0.041  0.6691254  0.3356520
   0.043  0.6691254  0.3356520
   0.045  0.6691254  0.3356520
   0.047  0.6691254  0.3356520
   0.049  0.6691254  0.3356520

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.007.
> |
```

Figure 4. Performance measure of decision tree

The Decision Tree Model performs superior than Logistic Regression. And the comparison of both the model based on different parameter are shown below.

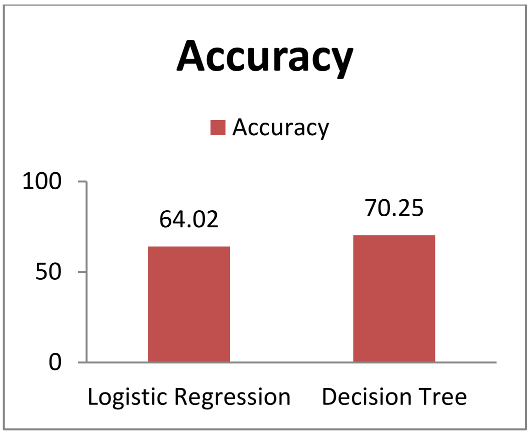


Figure 5. Accuracy Comparison

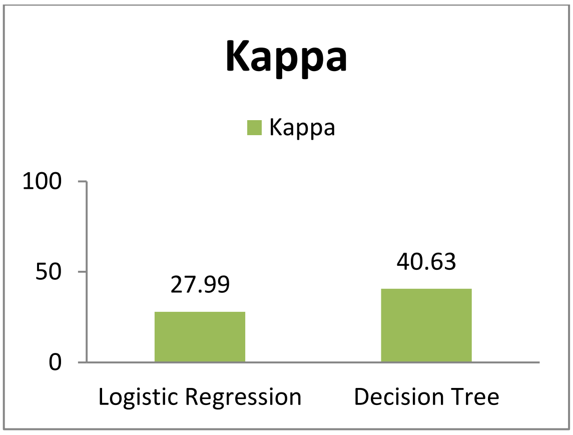


Figure 6. Kappa Comparison

VI CONCLUSION

Analysis of data which is extracted from telecom companies can helps to find the reasons of client churn and furthermore utilizes the data to hold the client. So predicting churn is very essential for telecom organizations to hold their client. From the result analysis we say that decision tree algorithm give better result as compared with the logistic regression model. By comparing the accuracy value based on the criteria, the decision tree Model gives better result than the logistic regression model (LRM).

REFERENCES

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

[02] Chuanqi Wang, Ruiqi Li, Peng Wang, Zonghai Chen, “Partition cost-sensitive CART based on customer value for Telecom customer churn prediction” in IEEE 2017.

[03] Duyen DO, Phuc HUYNH, Phuong VO, Tu VU, “Customer Churn Prediction in an Internet Service Provider” in 2017 IEEE.

[04] AMMAR A AHMED, Dr. D. Maheswari linen, “A Review And Analysis Of Churn Prediction Methods For Customer Retention In Telecom Industries” in 2017 International Conference on Advanced Computing and Communication Systems (ICACCS -2017), Jan. 06 – 07, 2017, Coimbatore, INDIA.

[05] Qiu Yihui. "Research of indicator system in customer churn prediction for telecom industry." in IEEE, 2016.

[06] Aimée Backiel and GerdaClaeskens. "Predicting time-to-churn of prepaid mobile telephone customers using social network analysis." In JORS, 2016.

[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] Ionuţ Brânduşoiu, HoriaBeleiu "Methods for churn prediction in the pre-paid mobile telecommunications industry." in IEEE, 2016.

[09] S. A. Qureshi, A. M. Qamar, A. Rehman “Telecommunication subscribers’ churn prediction model using machine learning,” in ICDIM, 2013.