Predicting Churn In Telecom Sector Using Classification & Decision Tree

Rohit jain
rohitjhankar@gmail.com
M.Tech(CSE)* ,LNCT, Bhopal

Prof.Vikram rajpoot
Rajputvikram022@gmail.com
Prof. Dept.ofCSE, LNCT,Bhopal

Dr.Vineet Richhariya
Vineetrich100@gmail.com
Prof.Dept.ofCSE,LNCT,Bhopal

Abstract— **Customer churn prediction in Telecom industry is one of the most prominent research topics in recent years. It consists of detecting customers who are likely to cancel a subscription to a service. Recently, the mobile telecommunication market has changed from a rapidly growing market into a state of saturation and fierce competition. The focus of telecommunication companies has therefore shifted from building a large customer base into keeping customers in house. For that reason, it is valuable to know which customers are likely to switch to a competitor in the near future. The data extracted from telecom industry can help analyze the reasons of customer churn and use that information to retain the customers. So churn prediction is very essential in telecom industries to retain their customers. In this paper we can use classification techniques along with decision tree to better predicting churn in telecom sector.**

Keywords—Churn prediction, data mining, telecom system ,Customer retention,decision tree, classification system.

I.INTRODUCTION

Customer churn refers to when a customer switches from one service provider to another [3]. Churn is a problem for any provider of a subscription service or recurring purchasable. The focus of this paper is mainly on telecom industry because of its tremendous growth in the recent years. With easy communication and a number of service providers almost everyone today has a telecom subscription. Churn is especially important to mobile phone service providers because it is easy for a subscriber to switch services. Phone number portability has removed the last important obstacle. Churn Prediction model can help analyze the historical data available with the business to find the list of customers which are at high risk to churn. This will help the telecom industry to focus on a specific group rather than using retention strategies on every customer. Individualized customer retention is difficult because businesses usually have a big customer base and cannot afford to spend much time and money for it. However, if we could predict in advance which customers are at risk of leaving, we can reduce customer retention efforts by directing them solely toward such customers.

This is where the churn prediction model [4] can help the business to identify such high risk customers and thereby helps in maintaining the existing customer base and increase in revenues. Churn prediction is also important because of the fact that acquiring new customers is much costly than retaining the existing one. As the telecom users are billions in number even a small fraction of churn leads to high loss of revenue. Retention has become crucial especially in the present situation because of the increasing number of service providers and the competition between them, where everyone is trying to attract new customers and lure them to switch to their service.

With a large customer base and the information available about them data mining techniques proves to be a viable option for making predictions about the customers that have high probability to churn based on the historical records available. The data mining techniques can help find the pattern among the already churned customers and provide useful insights which can then be used strategically to retain customers.

II. LITERATURE REVIEW

According to [1] ,In this competitive world, business is becoming highly saturated. Especially, the field of telecommunication faces complex challenges due to a number of vibrant competitive service providers. Therefore, it has become very difficult for them to retain existing customers. Since the cost of acquiring new customers is much higher than the cost of retaining the existing customers, it is the time for the telecom industries to take necessary steps to retain the customers to stabilize their market value. This paper explores the application of data mining techniques in predicting the likely churners and attribute selection on identifying the churn.

In this [1], we presented a churn prediction methodology for telecom consumers. The problem is modeled as a binary classification problem and hence several state-of-the-art classifiers are used for building churn model. Performances of these classifiers are compared based on four performance metrics using two real publicly available telecom datasets. Based on the experimental results, we observe that the Gradient boost classifier outperforms others. Further, the importance of each customer attribute is analyzed and found that only six attributes have more than 50% importance towards churn prediction.

In [2], Customer churn leads to the losses of enterprise. To deal with the customer churn problem of the customer relationship management, this paper set up the model based on the characteristics of amount and imbalance data and verify on the real data of telecom. By comparing with the Bayes, Decision Tree (DT), Artificial Neural Networks (ANN) and Support Vector Machine (SVM), the ensemble learning algorithms have the potential advantages. The effect of ensemble is obvious advantage especially the base classifiers are Support Vector Machines and has better hit rate, lift coefficient and accuracy rate. It can be used as an effective measure for customer churn prediction.

Customer churn research [5] is an important aspect of customer relationship management, based on the comparison of each algorithm and ensemble learning, theory and Practice, based on ensemble learning and selective ensemble learning is an effective means to predict customer churn, there are still a lot of problems, for example, how to choose the method of integration, how to choose the strategy, which makes the final ensemble classifier has the best generalization ability, and how to select the parameters of each algorithm as well as the kernel function calls to achieve the best results will be the focus of future research. On the whole, there is no absolute good classifier, in the face of different data using the appropriate classifier or classification method, can be classified forecast to do the best. When choosing a classifier, it is time to consider all aspects, horizontal and vertical contrast to find the most suitable, several classifiers are required to predict, based on the analysis of each focus, in order to get the most satisfactory results.

III PROBLEM DEFINITION

In a business environment, the term, customer attrition simply refers to the customers leaving one business service to another. Customer churn or subscriber churn is also similar to attrition, which is the process of customers switching from one service provider to another anonymously. From a machine learning perspective, churn prediction is a supervised (i.e. labeled) problem defined as follows: Given a predefined forecast horizon, the goal is to predict the future churners over that horizon, given the data associated with each subscriber in the network. The churn prediction problem represented here involves 3 phases, namely, i) training phase, ii) test phase, iii) prediction phase. The input for this problem includes the data on past calls for each mobile subscriber, together with all personal and business information that is maintained by the service provider. In addition, for the training phase, labels are provided in the form of a list of churners. After the model is trained with highest accuracy, the model must be able to predict the list of churners from the real dataset which does not include any churn label. In the perspective of knowledge discovery process, this problem is categorized as predictive mining or predictive modeling. Churn Prediction [6] is a phenomenon which is used to identify the possible churners in advance before they leave the network. This helps the CRM department to prevent subscribers who are likely to churn in future by taking the required retention policies to attract the likely churners and to retain them. Thereby, the potential loss of the company could be avoided.

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 model for churn prediction[7].

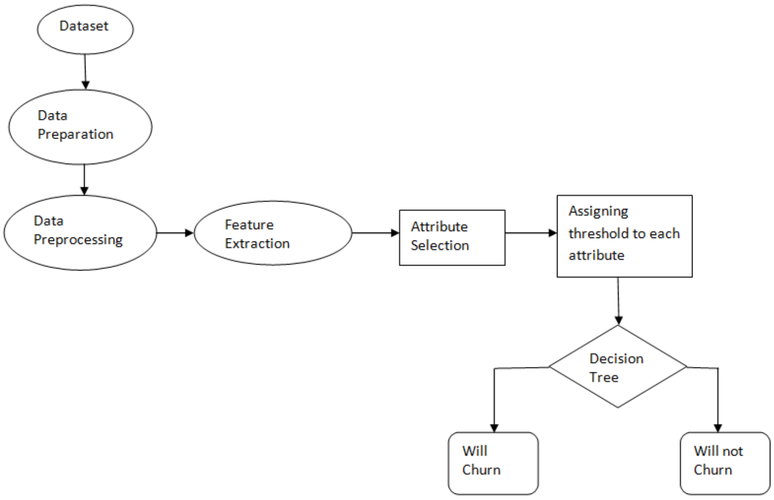


Figure1. Churn Prediction Framework

The system will have main options namely View performance analysis[9] – which displays the results obtained by applying classification and regression decision tree on the available dataset, Testing – to construct a list of customers which have a high probability to churn from the input, given that the attributes of the input data are same as the available dataset used for training, Training and testing – which builds a model along with generating a churn list if any other type of dataset is provided. In performance analysis the results after using logistic regression and decision trees on the available dataset is illustrated using confusion matrix analysis. In the next operation the user can provide data for testing the system provided the features of the data are same as that used for training using the publicly available dataset.

V. PROPOSED METHODOLOGY:

Our Steps or Algorithm Steps will follow:

1. Dataset:- we first download the training data

2. set from publicly available datasets.
3. Data Preparation: Since the dataset acquired cannot be applied directly to the churn prediction models, so aggregation of data is required where new variables are added to the existing variables by viewing the periodic usage behavior of the customers. These variables are very important in predicting the behavior of customers in advance as they contain critical information used by the prediction models.
4. Data Preprocessing: Data preprocessing is the most important phase in prediction models as the data consists of ambiguities, errors, redundancy which needs to be cleaned beforehand. The data gathered from multiple sources first is aggregated and then cleaned as the complete data collected is not suitable for modeling purposes. The records with unique values do not have any significance as they do not contribute much in predictive modeling. Fields with too many null values also need to be discarded.
5. Data Extraction: The attributes are identified for classifying process. In our work, we have worked with numerical and categorical values.
6. Decision: The rule set will let the subscribers identify and classify in the different categories of churners and non churners by setting a particular threshold
7. value.

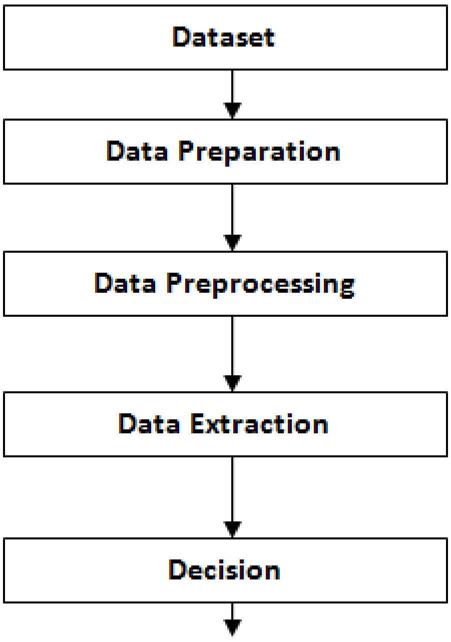
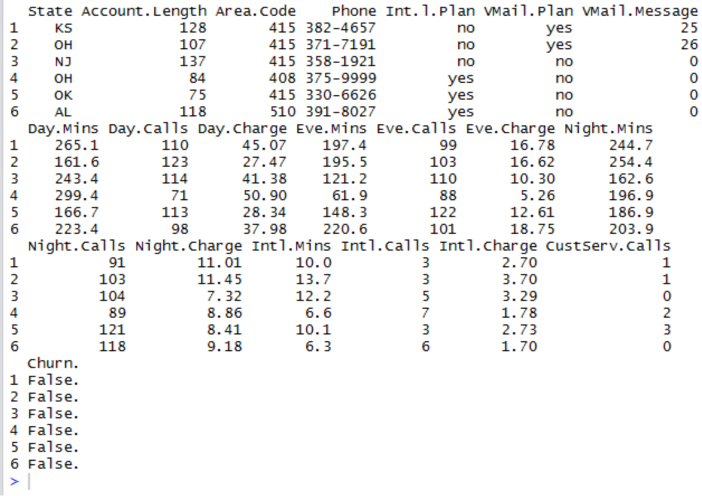


Figure 2. Analysis Steps

VII EXPERIMENTAL & RESULT ANALYSIS

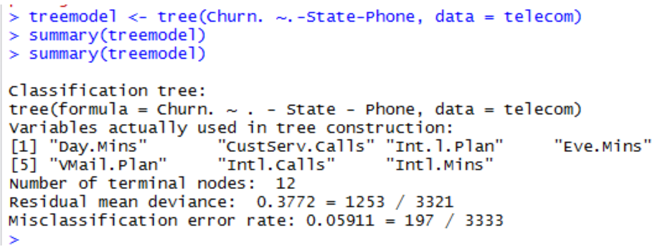
All the experiments were performed using an i5-2410M CPU @ 2.30 GHz processor and 4 GB of RAM running Windows. After that we can install r base core on windows and Rstudio and than to identify trends in customer churn at a telecom company. The data given to us contains 3,333 observations and 23 variables extracted from a data warehouse. These variables are shown in figure 3.



```
  State Account.Length Area.Code    Phone Int.l.Plan VMail.Plan VMail.Message
1    KS            128       415 382-4657         no        yes            25
2    OH            107       415 371-7191         no        yes            26
3    NJ            137       415 358-1921         no         no             0
4    OH             84       408 375-9999        yes         no             0
5    OK             75       415 330-6626        yes         no             0
6    AL            118       510 391-8027        yes         no             0
  Day.Mins Day.Calls Day.Charge Eve.Mins Eve.Calls Eve.Charge Night.Mins
1    265.1       110      45.07    197.4        99      16.78      244.7
2    161.6       123      27.47    195.5       103      16.62      254.4
3    243.4       114      41.38    121.2       110      10.30      162.6
4    299.4        71      50.90     61.9        88       5.26      196.9
5    166.7       113      28.34    148.3       122      12.61      186.9
6    223.4        98      37.98    220.6       101      18.75      203.9
  Night.Calls Night.Charge Intl.Mins Intl.Calls Intl.Charge CustServ.Calls
1          91        11.01      10.0          3        2.70              1
2         103        11.45      13.7          3        3.70              1
3         104         7.32      12.2          5        3.29              0
4          89         8.86       6.6          7        1.78              2
5         121         8.41      10.1          3        2.73              3
6         118         9.18       6.3          6        1.70              0
  Churn.
1 False.
2 False.
3 False.
4 False.
5 False.
6 False.
> 
```

Figure-3. Variables or sample values in datasets

Than we build our classification decision tree. Notice, that we specify -State and -Phone in the tree formula. This is because these categorical variables have too many levels. Running the tree with these variables included will result in an error message and so we remove them. Figure 4 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

The output tells that that R used 7 variables in the decision tree and that we have a total of 12 terminal nodes. We also see that the *residual mean deviance* (RMD) is 0.3372 an we have a ~5% misclassification rate. RMD is a metric that indicates how well the tree fits the data. In the real world however, we would be more interested in the RMD of the tree on the *test* dataset. Generally, a lower value indicates a better fitted decision tree.

Then we use the plot fuction and plot the classification decision tree which is shown in figure 5.

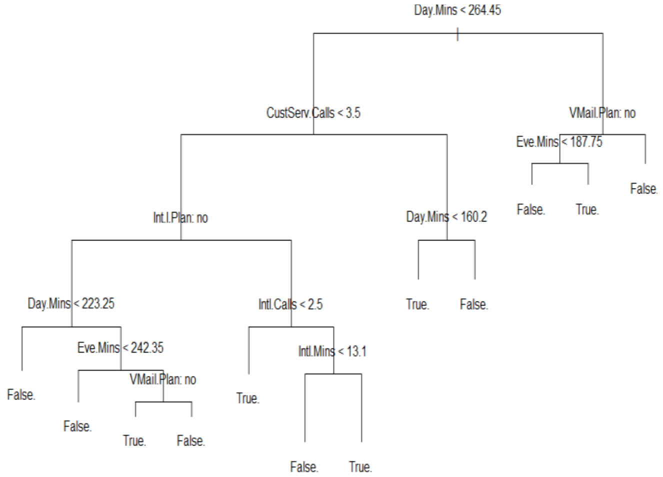
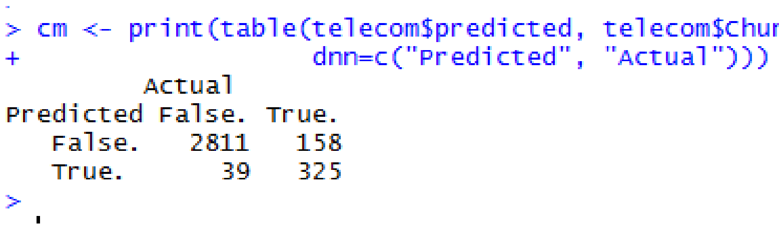


Figure-5. Decision tree

In the basis of decision tree we can predict the churn, and the output values predicted are shown in figure 6.



```
> cm <- print(table(telecom$predicted, telecom$Chur
+                         dnn=c("Predicted", "Actual"))
          Actual
Predicted False. True.
   False.   2811   158
   True.      39   325
>
```

Figure-6. Predicted output

In the output we can seen that there are predicted values comes along with actual value, means in that we predicted that customer cannot churn but 158 peoples are churn , and we predticed 325 peoples are churn. Now we can proposed a new decision tree which is based on pruning the classification decision tree which improves for detecting or predicting churn.

Now we can prune the tree with the help of miss.class function means whose terminal node having maximum misclassification erro we can prune that nodes and the new classification tree comes are shown in figure 7.

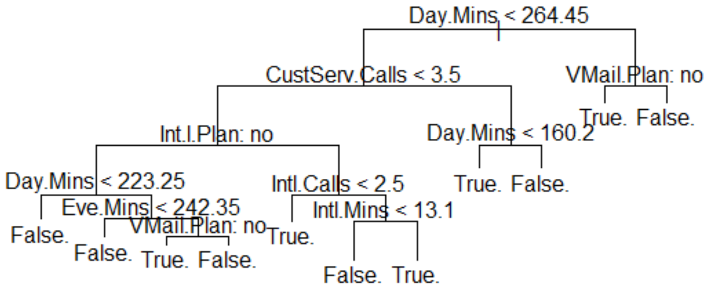
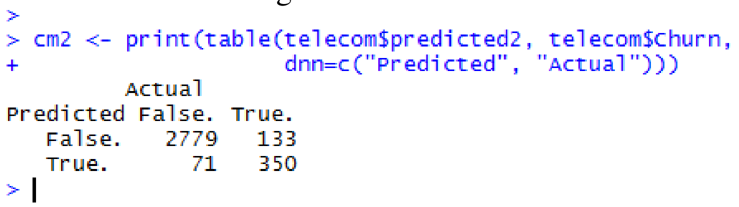


Figure-7. Decision tree after pruning

Now again we can calculating a predicting values which are shown in figure 8.



```
>
> cm2 <- print(table(telecom$predicted2, telecom$Churn,
+                    dnn=c("Predicted", "Actual"))
         Actual
Predicted False. True.
   False.   2779   133
   True.      71   350
> |
```

Figure-8. Proposed predicting output

When comparing these result into the above previous output values it is clear that we can the predicting churn is more accurate in proposed . the true positive rate known as the recall or probability of detection measures the proportion of positives that are correctly identified as churn.

VI CONCLUSION:

In this paper, it is observed that decision tree model better in the prediction of churn and it is also easy to construct. Selecting the right combination of attributes and fixing the proper threshold values may produce more accurate results. As we can see, pruning the tree to 11 nodes resulted in an improvement to sensitivity (the recall) at a slight cost to specificity and model accuracy.

REFERENCES

[01] V. Umayaparvathi, K. Iyakutti, “Attribute Selection and Customer Churn Prediction in Telecom Industry” in IEEE , december 2016, 10.1109/SAPIENCE.2016.7684171

[02] Guo-en Xia, Hui Wang, Yilin Jiang, “Application of Customer Churn Prediction Based on Weighted Selective Ensembles” in The 2016 3rd International Conference on Systems and Informatics (ICSAI 2016), IEEE 2016.

[03] Giridhar Maji, Soumya Sen, “Data Warehouse Based Analysis on CDR to Retain and Acquire Customers by Targeted Marketing” in 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), Sep. 7-9, 2016

[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.