PREDICTING CUSTOMER CHURN PREDICTION IN TELECOM SECTOR USING VARIOUS MACHINE LEARNING TECHNIQUES

Abhishek Gaur
abhisheklnct01@gmail.com

Ratnesh Dubey
Assistant Prof. of Dept.(CSE),
LNCT, Bhopal
ratneshdub@gmail.com

***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 various machine learning techniques for predicting customer churn through which we can build the classification models such as Logistic Regression, SVM, Random Forest and Gradient boosted tree and also compare the performance of these models.**

Keywords—Churn prediction, data mining, telecom system ,Customer retention, classification system, random forest, logistic regression.

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 lost 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 the paper [1] , From the beginning of the data mining [9] 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 In these paper they 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[10]. In this they are using logistic regression model 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 so we can get the various parameters like true positive rate, false positive rate and accuracy.

According to [2] , Telecom Customer churn prediction is a cost sensitive classification problem. Most of studies regard it as a general classification problem use traditional methods, that the two types of misclassification cost are equal. And, in aspect of cost sensitive classification, there are some researches focused on static cost sensitive situation. In fact, customer value of each customer is different, so misclassification cost of each sample is different. For this problem, we propose the partition cost-sensitive CART model in this paper[12]. According to the experiment based on the real data, it is showed that the method not only obtains a good classification performance, but also reduces the total misclassification costs effectively.

978-1-5386-5367-8/18/$31.00 ©2018 IEEE

According to paper [5] Customer churn plays an important role in customer relationship management (CRM), and they are using various machine learning algorithm to predict customer churn and they found ensemble learning is an best to predict customer churn, but there exist still a lot of problems like how they choose the method of integration and how to choose the strategy, which makes the final ensemble classifier. On the other hand, there is no good classifier, so its also a main problem to chosen which classification algorithm is best for which situation. So we can consider various aspect like vertical and horizontal contrast to find the best classifier to predict the customer churn in telecom sector.

III.PROBLEM DEFINITION

From the problems obligatory through market saturation and value implications, there has been associate identification of a desire for a laptop 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. These offers is pricey, thus providing them to customers World Health Organization don't shall churn will have a substantial impact on the whole price of a retention strategy. an extra limitation of current analysis is that it's typically supported one output within the kind of zero for non-churn and one for churn. This has been recognized as a limitation as a result of it restricts analysis prospects.

So as to handle the issues mentioned higher than, a profile based mostly analysis methodology is known as a doable answer. it's anticipated that profile based mostly analysis can change future prediction, through the flexibility to match customers to profile clusters that are known as most fitted for capturing future churn. it's conjointly anticipated that profile analysis can offer a technique for dominant misclassification levels through eliminating the profile clusters that statistically hold the tiniest future churn capture accuracy.

Client churn management [13] is viewed from 2 separate angles. One space of client churn management focuses on churn interference. This includes investigations like contestant analysis, rating and repair ways. the world being addressed by this analysis is churn prediction [14]. during this space, it's common to use demographic and usage knowledge as predictor variables wherever client churn is that the response variable. This analysis identifies another approach to the utilization of demographic and usage knowledge by showing however repairs and complaints knowledge is used with success for the prediction of client churn.

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].

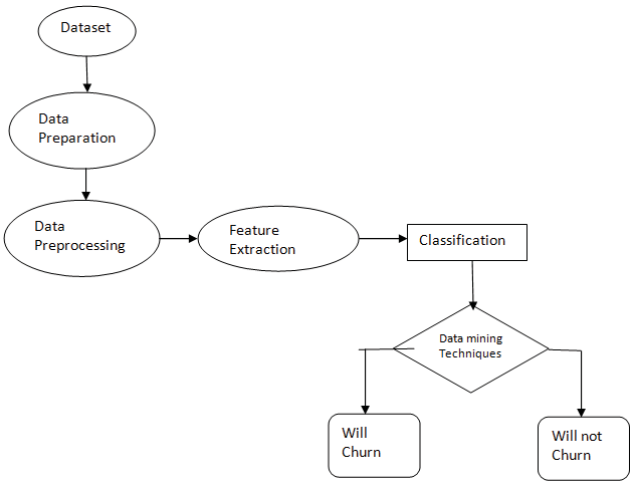

Figure 1. Churn Prediction Framework

In this paper, we proposed different machine learning 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 Logistic Regression[11], Support Vector Machine, Random Forest, Gradient Boosting Trees.

Our Steps or Algorithm Steps will follow:

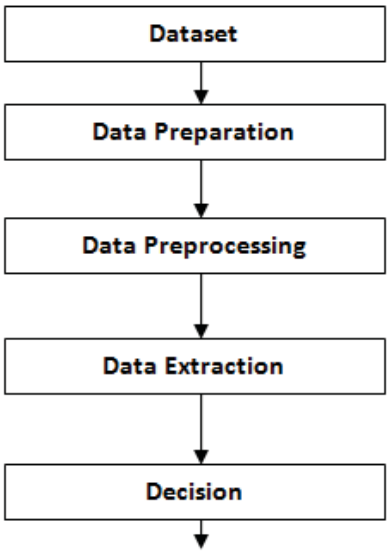

Figure 2. Analysis Steps

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

978-1-5386-5367-8/18/$31.00 ©2018 IEEE

data given to us contains 7043 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 and Rstudio and than to identify trends in customer churn at a telecom company. The data given to us contains 7043 observations and 21 variables extracted from a data warehouse. These variables are shown in figure 3.

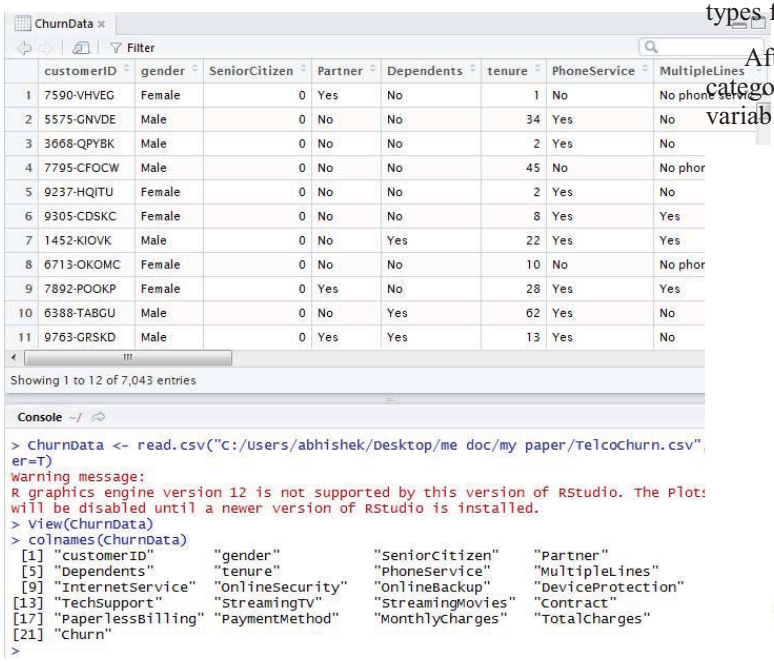

Figure-3. Variables or sample values in datasets

Now we started to exploring a data and cleaning a data for machine learning models, we can explore the data by their multiple attributes such as Average monthly charge of those who churned, Average total charge of those who churned and also visualize these exploring result such as Average monthly charges by internet service types for churned customer which is shown in figure 4.

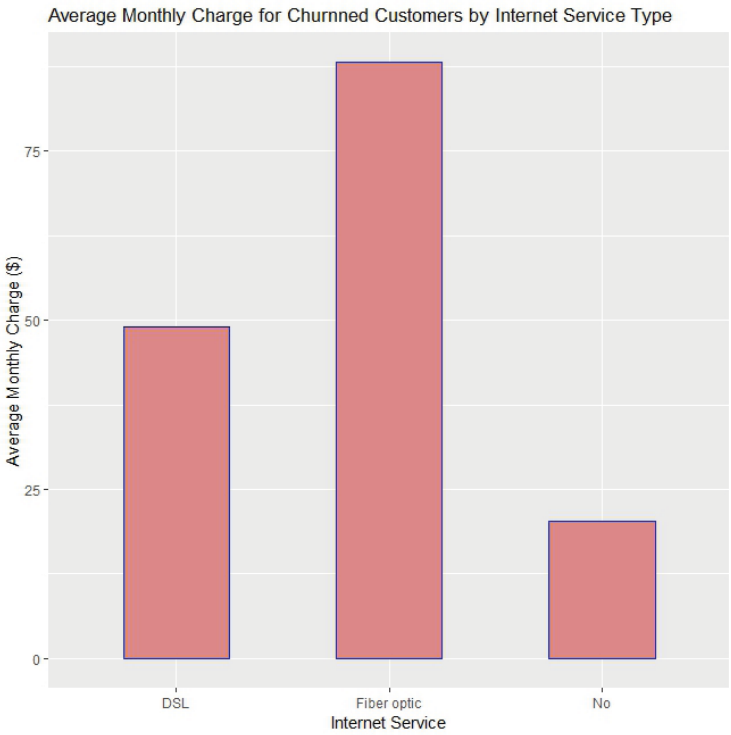

Figure 4. Average monthly charges by internet service types for churned customer

After that we can find the Correlation between these categorical variables figure 5 shows the relation between these variables.

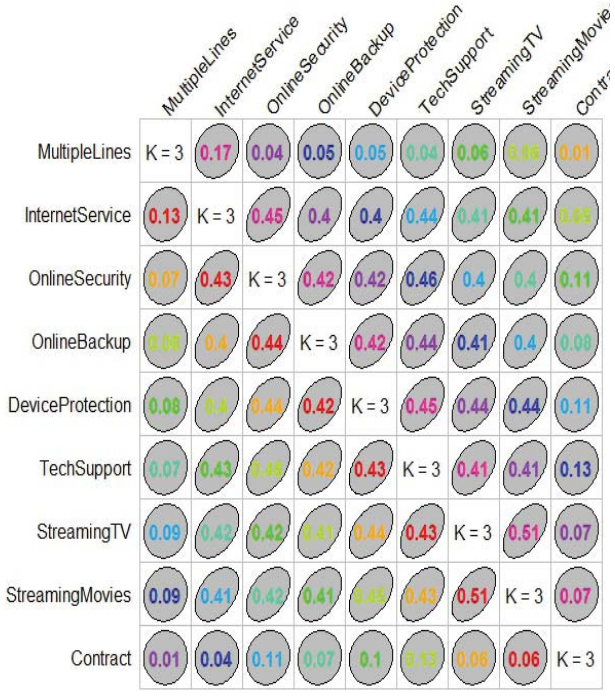

Figure 5. correlation between categorical variables

978-1-5386-5367-8/18/$31.00 ©2018 IEEE

These graph is to explore relationships between categorical variables and which look like the only highly correlated variables are "streaming movies" and "streaming tv" which is expected.

PREPARATION FOR THE MODEL BUILDING

Now we can build a various machine learning models, such as Logistic Regression, SVM, Random Forest and Gradient boosted tree and then we train these classifier and after training we can test these model and compare their performance. Before starting training we can perform variable selection: We know that we should not include "streaming movies" and "streaming tv" in the same equation. Their correlation from the above section is fairly high.

Now we can split the data into train (75%) and test (25%) dataset and then start learning of model on these train data and we can test the model on test dataset and compute model measures. Figure 6 shows the compting performance of these four machine learning models.

```
Console ~/
> ROCRperfGBM <- performance(ROCRpredGBM, "tpr", "fpr")
> plot(ROCRperfGBM)
> aucGBM <- performance(ROCRpredGBM,measure='auc')
> aucGBM <- aucGBM@y.values[[1]]
> aucGBM
[1] 0.8459888
> m <- matrix(c(auc,aucSVM,aucwholeRandom,aucGBM),nrow=4,ncol=1)
> colnames(m) <- c("AUC Value")
> rownames(m) <- c("Logistic Regression","SVM","Random Forest","Gradie
> m
                   AUC Value
Logistic Regression 0.8286128
SVM                 0.7975093
Random Forest       0.8126890
Gradient Boosting   0.8459888
>
```

Figure 6. AUC values of models

According to the AUC values which we have computed, the method that gives us the most accurate model is gradient boosting with AUC value of 84.57%.

ROC curve shows the tradeoff between sensitivity and specificity: The measure of **sensitivity** is the proportion of positive examples that were correctly classified. The measure of **specificity** is the proportion of negative examples that were correctly classified.

AUC: The area under the curve. It is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. Figure 7 show the AUC curve of these four models.

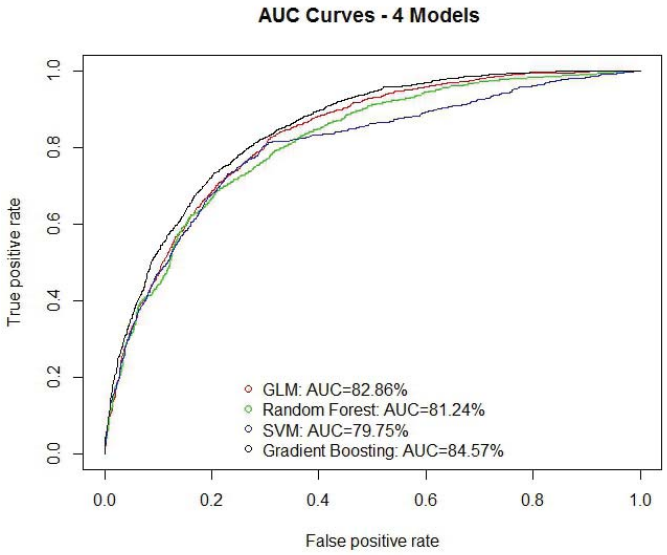

Figure 7. AUC curves of four models

According to the AUC curves, the method that gives us the most accurate model is gradient boosting with AUC value of 84.57%. And based on the AUC curve we can compare the performance of the models and the conclusion of the performance are shown in figure 8.

```
Console ~/
> aucTestGBM <- performance(ROCRpredTestGBM,measure= 'auc')
> aucTestGBM <- aucTestGBM@y.values[[1]]
> aucTestGBM
[1] 0.858874
> a <- matrix(c("The Best Model","Average Model","Average Model","Underperforming"),nrow
=4,ncol=1)
> colnames(a) <- c("General Performance (accuracy) of the algorithms")
> rownames(a) <- c("Gradient Boosting","Logistic Regression","Random Forest","Support Ve
ctor Machine")
> a
                     General Performance (accuracy) of the algorithms
Gradient Boosting    "The Best Model"
Logistic Regression  "Average Model"
Random Forest        "Average Model"
Support Vector Machine "Underperforming"
>
```

Figure 8. Performance on the models

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 train four machine learning models which is Logistic Regression, SVM, Random Forest and Gradient boosted tree and we can say that Gradient boosting is best in among four models and the Logistic regression and Random forest is an average and SVM is underperforming between these models.

978-1-5386-5367-8/18/$31.00 ©2018 IEEE

REFERENCES

[01] *Peng Li 1, 2, Siben Li 2, Tingting Bi 2, Yang Liu 2, "* Telecom Customer Churn Prediction Method Based on Cluster Stratified Sampling Logistic Regression*" in IEEE*.

[02] Chuanqi Wang, Ruiqi Li, Peng Wang, Zonghai Chen, "Partition cost-sensitive CART based on customer value for Telecom customer churn prediction" in Proceedings of the 36th Chinese Control Conference 2017 IEEE.

[03] Guo-en Xia, Hui Wang, Yilin Jiang, "Application of Customer Churn Prediction Based on Weighted Selective Ensembles" in IEEE 2016.

[04] Rahul J. Jadhav, Usharani T. Pawar, "Churn Prediction in Telecommunication Using Data Mining Technology", in *(IJACSA), 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 *IJCA Volume 64– No.5, February 2013*

[07] Kiran Dahiya,KanikaTalwar, "Customer Churn Prediction in Telecommunication Industries using Data Mining Techniques- A Review" in IJARCSSE, 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] Manjit Kaur et al., 2013.Data Mining as a tool to Predict the Churn Behaviour among Indian bank customers, IJRITCC, Volume: 1 Issue: 9

[11] R. Khare, D. Kaloya, C. K. Choudhary, and G. Gupta, "Employee attrition risk assessment using logistic regression analysis,".

[12] Praveen et al., Churn Prediction in Telecom Industry Using R, in (IJETR) ISSN: 2321-0869, Volume-3, Issue-5, May 2015

[13] J. Burez and D. Van den Poel, "Handling class imbalance in customer churn prediction," *Expert Systems with Applications*, vol. 36, no. 3, 2009.

[14] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach," *European Journal of Operational Research*, vol. 218, no. 1, pp. 211–229, 2012.

978-1-5386-5367-8/18/$31.00 ©2018 IEEE