International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 2 Feb 2022, pp: 1446-1453 www.ijaem.net ISSN: 2395-5252

# Predicting Customer Churn in Telecommunication Sector Using Machine Learning

Manisha Srivastava, Purnendu Shekhar Pandey
*M.tech Scholar Department of Computer science and Engg.*
*Associate Professor Department of Electronics & Communication Engg.*

Submitted: 15-02-2022 Revised: 25-02-2022 Accepted: 28-02-2022

**ABSTRACT**— Predicting customer churn in telecommunication industries becomes a most important topic for research in recent years. Because its helps in detecting which customer are likely to change or cancel their subscription to a service. Now a days the mobile telecom market has growing market rapidly and all the telecommunication industries focused on building a large customer base into keeping customers in house. So it is very important to find which customers are wants to switch to a other competitor by cancel their subscription in the near future. 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. 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, We will built Logistic Regression, Decision Tree, Random Forest and XgBoost models and compare them and we can say that XgBoost and logistic Regression is perform better as compared to Decision Tress and Random Forest because it provides better accuracy.

**KEYWORDS—**Churn prediction, data mining, telecom system ,Customer retention, classification system.

## I.INTRODUCTION

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 thesis we can use classification techniques along with decision tree to better predicting churn in telecom sector.

Predicting customer churn in telecommunication industries becomes a most important topic for research in recent years. Because its helps in detecting which customer are likely to change or cancel their subscription to a service. Now a days the mobile telecom market has growing market rapidly and all the telecommunication industries focused on building a large customer base into keeping customers in house. So it is very important to find which customers are wants to switch to a other competitor by cancel their subscription in the near future. 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. In this we can focuses on various data mining techniques for predicting customer churn.

## II. LITERATURE REVIEW

According to the paper [1] Nowadays data has become the important aspect in each and every field. In this the data about the telecommunication industry is collected and then the raw data is classified into churn and the non churn customers. The churn customers are one who periodically uses the same resource signals and non churn customers are one who utilizes the resources based on the services provided by the particular company. In existing system they uses the algorithm called LDT and UDT which train the system blindly with too

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 2 Feb 2022, pp: 1446-1453 www.ijaem.net ISSN: 2395-5252

many attributes which are not necessary for the computation. So it takes much time to train the system and the accuracy is not that much efficient and it achieve the performance about 84 percent. But this much of performance is not that much efficient for an organization to provide convincible services. So in order to resolve this problem in existing system we proposing the system with an efficient algorithms known as Random Forest Classifier and Support Vector Machine which selects the important attribute which increases the performance of the system and by implementing these two algorithms we can achieve the efficiency of about 95 percent. Because this efficiency in performance will ensure the company to provide the appropriate services to retain the non churn customer within the organization to sustain the Telecommunication industry.

According to (Manpreet Singh, et al., 2018) [02] ,Customer Churn is a challenging and one of the most demanding issues in the telecom sector. The primary motivation of businesses at present is just not only to acquire new customers, but to retain existing customers as well. In fact, customer retention is more important because of the associated high costs. The present work has been carried out in a churn prediction modeling context and benchmarks four machine learning techniques against a publicly available telecommunication dataset. The results provide two important conclusions: i) Random Forest technique outperforms other basic classification models and ii) Feature Engineering plays critical role in the performance of the model.

Customer churning refers to the migration of a customer from one organization to another. Customer churns are those targeted customers that have already decided to leave the company or the service provider and planned to shift to the competitor's company in the market. Customer churn is one of the rapidly growing issue in the telecom sector. The high cost involved in acquiring a new customers has resulted inthe change of focus of telecom sector from acquiring new customers to retaining new customers.

According to (Muhammad Ali ,et al, 2018) [03], Data mining is vast area that co-relates diverse branches i.e Statistics, Data Base, Machine learning and Artificial intelligence. Various applications are accessible in various areas. Churning of the Customer is the behavior when client never again needs to stay with his association with the company. Customer Churn Management is assuming essential job in client management. Nowadays different telecommunication companies are concentrating on distinguishing high esteemed and potential churning clients to expand benefit and share market. It is comprehended that making new clients are costlier than to holding existing client. There is a current issue that customer leave the organization because of obscure reasons. In our investigation, we predict churn behavior of the client by utilizing diverse data mining methods. It will in the long run help in breaking down client's behavior and characterize whether it is a churning client or not. We utilize online accessible data set available at Kaggle repository and for forecasting of Customer behavior we utilized different algorithms while we achieved 99.8% accuracy level using Bagging Algorithms.

## III PROBLEM DEFINITION

In a business setting, the term, client attrition merely refers to the purchasers exploit one business service to a different. client churn or subscriber churn is additionally kind of like attrition, that is that the method of shoppers shift from one service supplier to a different anonymously. From a machine learning perspective, churn prediction could be a supervised (i.e. labeled) downside outlined as follows: Given a predefined forecast horizon, the goal is to predict the longer term churners over that horizon, given the info related to every subscriber within the network. The churn prediction downside diagrammatical here involves three phases, namely, i) the training part, ii) testing part, iii) prediction section. The input for this downside includes the info on past necessitate every mobile subscriber, along with all personal and business data that's maintained by the service supplier. additionally, for the training section, labels are provided within the type of an inventory of churners. when the model is trained with highest accuracy, the model should be able to predict the list of churners from the important dataset that doesn't embody any churn label. within the perspective of information discovery method, this downside is categorized as prognostic mining or prognostic modeling.

## 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].

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 2 Feb 2022, pp: 1446-1453 www.ijaem.net ISSN: 2395-5252

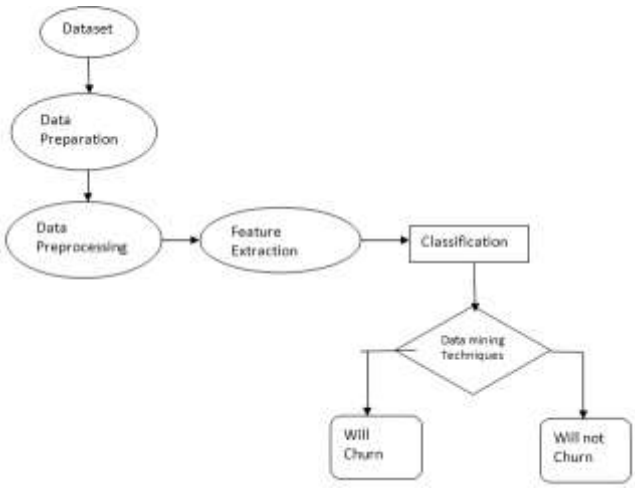


Figure 1. Churn Prediction Framework

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.
Our Steps or Algorithm Steps will follow:

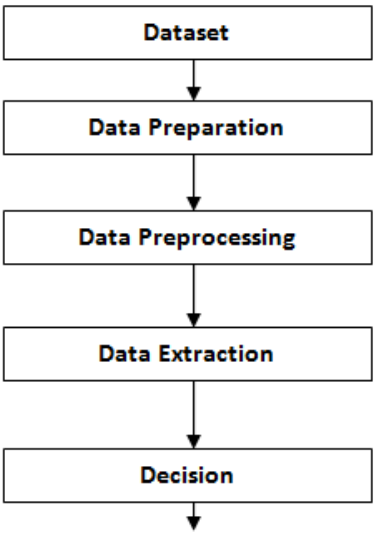


Figure 2. Analysis Steps

1. Dataset:- we first download the training data set from publicly available datasets.
2. Data Preparation: Since the dataset acquired cannot be applied directly to the churn prediction models, so aggregation of data is

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 2 Feb 2022, pp: 1446-1453 www.ijaem.net ISSN: 2395-5252

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.

3. 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.
4. Data Extraction: The attributes are identified for classifying process. we have worked with numerical and categorical values.
5. 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 value.

## V EXPERIMENTAL 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 then 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.

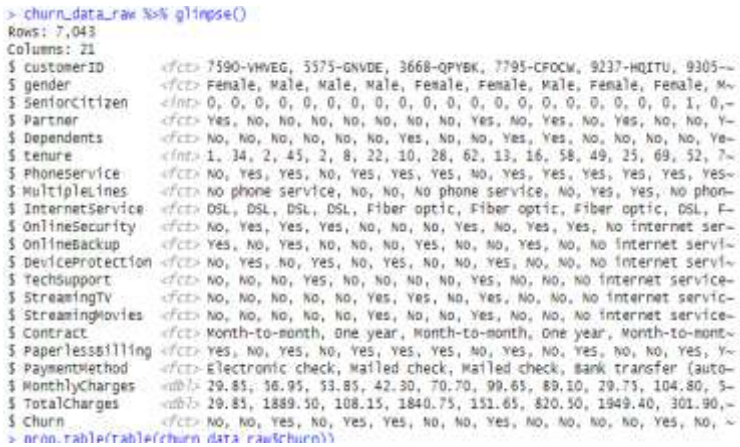


Figure-3. Variables or sample values in datasets

Now we started to preprocess the data and cleaning a data for machine learning models and check missing data, Remove the customerID variable as it a unique value and does not give any information, convert character features to factors and drop the NAs in TotalCharges as it is very small amount of data. And after these we can split thedataset into training and testing data. After that can start exploring the data, By Visualize numerical features with histogram shown in figure 4.

.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 2 Feb 2022, pp: 1446-1453 www.ijaem.net ISSN: 2395-5252

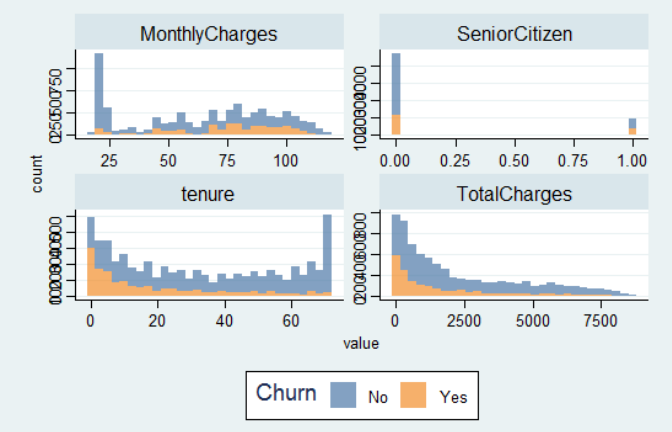


Figure 4 , Visualize numerical features with histogram

In these we seen that, SeniorCitizen seems to be categorical feature and so will be transformed to factor TotalCharges is right skewed and so will be transformed using log transformation, tenure could be discretized. Now Exploring categorical variables by Visualize the relation between categorical and target feature

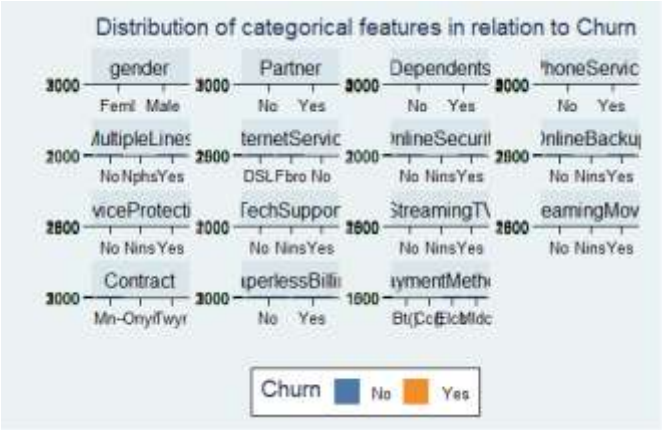


Figure 5 Visualize the relation between categorical and target feature

**Modeling**

Now after spliting the dataset we can start building the model, We will built Logistic Regression, Decision Tree, Random Forest and XgBoost models and compare them on training and testing dataset and the summary areshown in figure 6.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 2 Feb 2022, pp: 1446-1453 www.ijaem.net ISSN: 2395-5252

```
> summary(model_list)

Call:
summary.resamples(object = model_list)

Models: Logistic, Decision_Tree, Random_forest, XgBoost
Number of resamples: 10

ROC
                   Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
Logistic      0.8353186 0.8416201 0.8487373 0.8488915 0.8560513 0.8630417    0
Decision_Tree 0.7504428 0.7764211 0.8025222 0.7972051 0.8150666 0.8361961    0
Random_forest 0.8177357 0.8234725 0.8252880 0.8343312 0.8422821 0.8698913    0
XgBoost       0.8337127 0.8401830 0.8470067 0.8479512 0.8553793 0.8650649    0

Sens
                   Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
Logistic      0.8886199 0.8916465 0.9007264 0.9055690 0.9207022 0.9297821    0
Decision_Tree 0.8837772 0.8958838 0.9055690 0.9077482 0.9194915 0.9322034    0
Random_forest 0.9322034 0.9394673 0.9479419 0.9479419 0.9546005 0.9709443    0
XgBoost       0.8789346 0.8886199 0.9031477 0.9029056 0.9158596 0.9249395    0

Spec
                   Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
Logistic      0.4765101 0.5225503 0.5418121 0.5337852 0.5466667 0.5637584    0
Decision_Tree 0.3933333 0.4522483 0.4648770 0.4776376 0.5166667 0.5704698    0
Random_forest 0.3288591 0.3422483 0.3533333 0.3545324 0.3589262 0.3959732    0
XgBoost       0.5100671 0.5150000 0.5250783 0.5331275 0.5400000 0.5906040    0

>
```

Figure 6. Summary of classification model

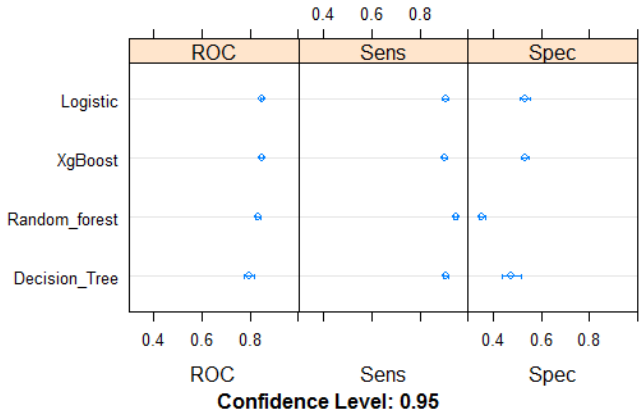


Figure 7. Model Comparison

Logistic regression and XgBoost is performing good with auc score as good as compared to Decision and Random forest with lower variance, After that we can test these models on testing dataset and the result are shown in figure 8 and Visualize ROC curves and gain curve which is shown in figure 9 and figure 10.

```
## # A tibble: 4 x 4
##   model    .metric .estimator .estimate
##   <fct>    <chr>   <chr>          <dbl>
## 1 Logistic roc_auc binary         0.842
## 2 Tree     roc_auc binary         0.772
## 3 RF       roc_auc binary         0.820
## 4 xgb      roc_auc binary         0.843
```

Figure 8. AUC comparison on Testing dataset

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 2 Feb 2022, pp: 1446-1453 www.ijaem.net ISSN: 2395-5252

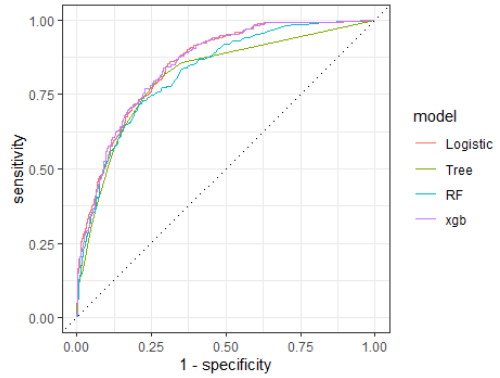


Figure 9. AUC Curve

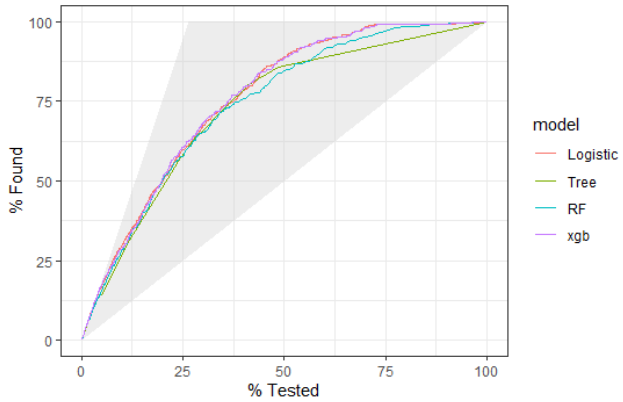


Figure 10. Gain Curve

From the gain curve we can see that by targeting about 37% of the potential churners we can correctly identify about 75% of the actual churners if we apply Logistic regression and xgb model.

Based on the AUC curve and the performance measures of the models we can compare the performance of the models and the conclusion of the performance are shown in figure 11.

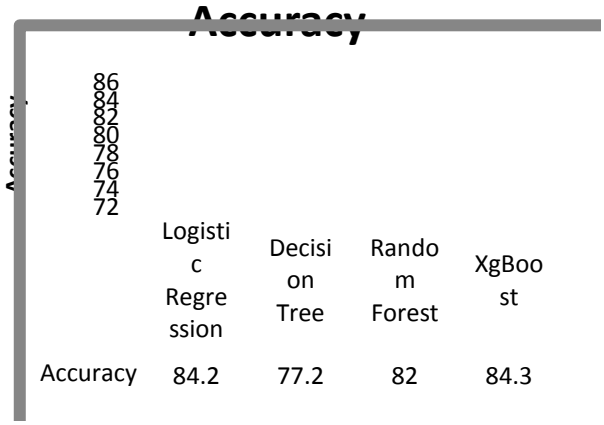


Figure 11. Performance on the models

## 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. 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, We will built Logistic Regression, Decision Tree, Random Forest and XgBoost models and compare

them and we can say that XgBoost and logistic Regression is perform better as compared to Decision Tress and Random Forest because it provides better accuracy .

## REFERENCES

[1]. V. Geetha; A. Punitha; A. Nandhini; T. Nandhini; S. Shakila; R. Sushmitha, "Customer Churn Prediction In Telecommunication Industry Using Random Forest Classifier, in IEEE 2020.
[2]. Manpreet Singh, Sarbjeet Singh, Nadesh Seen, Sakshi Kaushal and Harish Kumar, "Comparison of learning techniques for prediction ofcustomer churn in telecommunication" in IEEE 2018.
[3]. Muhammad Ali, Aziz Ur Rehman, Shamaz Hafeez, "Prediction of Churning Behavior of Customers inTelecom Sector Using Supervised LearningTechniques" in 2018 IEEE.
[4]. Rahul J. Jadhav, Usharani T. Pawar, "Churn Prediction in Telecommunication Using Data Mining Technology", in (IJACSA), Vol. 2, No.2, February 2011
[5]. Kiran Dahiya, Surbhi Bhatia, "Customer Churn Analysis in Telecom Industry" in IEEE 2015, 978-1-4673-7231-2/15
[6]. N.Kamalraj, A.Malathi' " A Survey on Churn Prediction Techniques in Communication Sector" in IJCA Volume 64– No.5, February 2013
[7]. Kiran Dahiya,KanikaTalwar, "Customer Churn Prediction in Telecommunication Industries using Data Mining Techniques- A Review" in IJARCSSE, Volume 5, Issue 4, 2015.
[8]. R Data: http://cran.r-project.org/
[9]. 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.