IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12744

CUSTOMER CHURN PREDICTION ON TELECOM DATA USING SUPERVISED MACHINE LEARNING ALGORITHMS

Surendra Singh[1], Sushma Kushwaha[2]

M.Tech (CSE), SVCST, Bhopal (M.P)[1]

Asst. Professor, SVCST, Bhopal (M.P)[2]

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 can focuses on machine learning techniques for predicting customer churn through which we can build the classification models such as logistic Regression, Random Forest and Gradient Boosting Algorithm and also compare the performance of these models.

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] Currently data has come the important aspect in each and every field. In this the data about the telecommunication assiduity is collected and also the raw data is classified into churn and the non churn guests. The churn guests are one who periodically uses the same resource signals and non churn guests are one who utilizes the coffers grounded on the services handed by the particular company. In being system they uses the algorithm called LDT and UDT which train the system blindly with too numerous attributes which aren't necessary for the calculation. So it takes important time to train the system and the delicacy isn't that important effective and it achieve the performance about 84 percent. But this important of performance isn't that important effective for an association to give convincible services.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023
DOI: 10.17148/IJARCCE.2023.12744

So in order to resolve this problem in being system we proposing the system with an effective algorithms known as Random Forest Classifier and Support Vector Machine which selects the important trait which increases the performance of the system and by enforcing these two algorithms we can achieve the effectiveness of about 95 percent. Because this effectiveness in performance will insure the company to give the applicable services to retain the non churn client within the association to sustain the Telecommunication assiduity.

According to [02] , Client churn vaticination is classic task of machine literacy, the applicability of which continues to grow. This is due to the fact that business companies collect further data about their guests and their geste every time. A model that predicts whether a client churn will do in the future allows a business to make an optimal substantiated pricing policy to retain a client. Being approaches for working the problem of churn vaticination for different areas are anatomized. A set of data from the Prozorro system was chosen for practical operation of the approaches.

According to [03], This paper presents churn prognostications with the Gaussian Na ¨ ıve Bayes system. Churn vaticination is a soothsaying system to prognosticate client opinions in a company's service or product (churn). With high public enthusiasm and an adding number of guests in the Big Data period, a fast computing process is demanded to prognosticate churn as snappily as possible.

In this paper, computing is accelerated by the OpenMP platform resemblant algorithm. Churn vaticination trials are performed with different quantities of test data, ranging from 100, 300, 500, 700, to 900 data. The results attained show that enforcing OpenMP in prognosticating churn is faster than periodical processing.

III. PROBLEM DEFINITION

To prognosticate the telecom guests who are likely to exit the contract and also to induce patterns of Churn andnon-churn to help the operation to take applicable opinions to limit churn. Utmost telecom companies suffer from voluntary churn. Churn rate has strong impact on the life time value of the client because it affects the length of service and the unborn profit of the company. It's estimated that 75 percent of the 17 to 20 million subscribers subscribing up with a new wireless carrier every time are coming from another wireless provider, which means they're churners.

Ways to Reduce Client Churn

- Constantly Exceed Guests' Prospects The most abecedarian way to drop your churn rate is by keeping your guests happy. While you surely want to avoid letting them down, you have to look for areas to go over and above your client's prospects and delight them. Failing to deliver on a pledge is one of the fastest ways to lose a client, and numerous companies say that dissatisfaction and unmet prospects are among the top reasons for customer churn.

- Give Stupendous Client Service This bone should go without saying, but if you 've ever spent half an hour harkening to hold muzak staying for a disinterested, unskillful client service rep to " help you," you 'll know that some companies simply do n't put enough trouble into client service.

- Produce Switching Costs Switching costs are any cost that a client incurs by trading one product or service for another. Advanced switching costs naturally reduce churn by reducing the liability that a client will switch to a cover product rather of returning to your brand.

IV. PROPOSED WORK

In the proposed system Python programming (Jupyter Notebook) will be used to make the model for churn vaticination. It's extensively used among statisticians and data miners for developing statistical software and data analysis. Jupyter Notebook is freely available and a important statistical analysis tool which has not yet been explored for structure models for churn vaticination.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12744

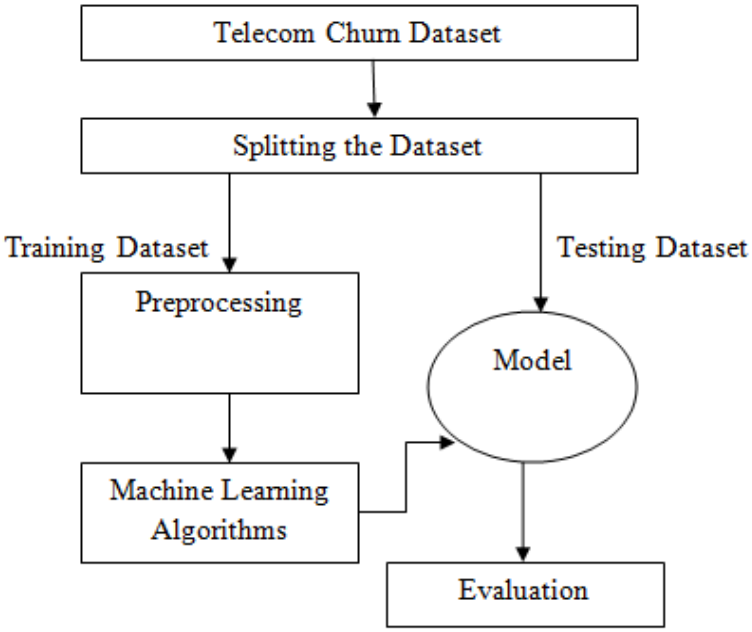

Figure 1. Proposed Churn Prediction Framework

This is wherever the churn vaticination model will grease the business to spot similar high threat guests and thereby helps in maintaining the prevailing client base and increase in earnings. Churn vaticination is also necessary thanks to the very fact that deed new guests is far precious than forgetful the prevailing one. because the telecommunication druggies area unit billions in variety indeed alittle bit of churn ends up in high loss of profit. Retention has come pivotal particularly within the gift state of affairs thanks to the adding variety of service suppliers and thus the competition between them, wherever everyone seems to be trying to draw in new guests and bait them to change to their service. \

With an outsized client base and therefore the info offered concerning them data processing techniques proves to be a viable possibility for creating predictions concerning the shoppers that have high chance to churn supported the historical records offered. the info mining techniques will facilitate notice the pattern among the already churned customers and supply helpful insights which might then be used strategically to retain customers.
Our Steps or Algorithm Steps will follow:

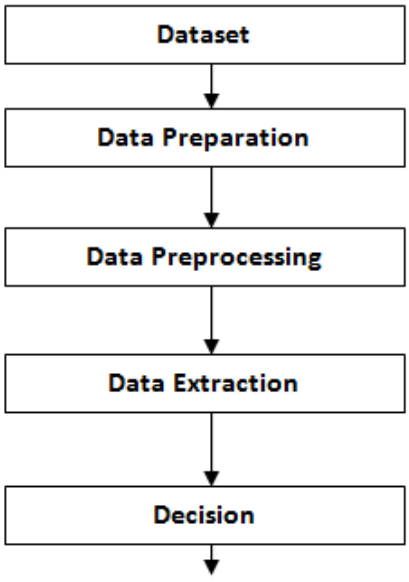

Figure 2. Analysis Steps

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12744

In this, we proposed different machine learning algorithms to dissect client churn analysis. Through which we can multiple different models are employed to directly prognosticate those churn guests in the data set. These models are Random Forest, XGBoost and Logistic Regression. Our way or Algorithm way will follow

1. Dataset- A telecom dataset is taken for prognosticating churn which to identify trends in client churn at a telecom company and the data which we taken is in. csv format.
2. Data Preparation Since the dataset acquired can not be applied directly to the churn vaticination models, so we can naming each attributes.
3. Data Preprocessing Data preprocessing is the most important phase in vaticination models as the data consists of inscrutability, crimes, redundancy and metamorphosis which needs to be gutted beforehand.
4. Data birth The attributes are linked for classifying process.
5. Decision Grounded on data birth and bracket models we can take a decision whether the hand is churner or not.

V. EXPERIMENTAL ANALYSIS

All the trials we are performing on intel i5 processor with 4 GigaBytes of RAM configured Windows Machine. For Experimental analysis purpose we uses python programming and jupyter notebook through which we can build our machine learning model. After launching jupyter notebook we can load the dataset and attribute are shown in figure 3.

```
df_load.head()
```

atedAt	customerID	gender	SeniorCitizen	Partner	tenure	PhoneService	StreamingTV	InternetService	PaperlessBilling	MonthlyCharges	TotalCharges	Churn
202006	45759018157	Female	No	Yes	1	No	No	Yes	Yes	29.85	29.85	No
202006	45315483266	Male	No	Yes	60	Yes	No	No	Yes	20.50	1198.80	No
202006	45236961615	Male	No	No	5	Yes	Yes	Yes	No	104.10	541.90	Yes
202006	45929827382	Female	No	Yes	72	Yes	Yes	Yes	Yes	115.50	8312.75	No
202006	45305082233	Female	No	Yes	56	Yes	Yes	Yes	No	81.25	4620.40	No

Figure-3. Variables or sample values in datasets

Exploratory Data Analysis (EDA)

Once the dataset have been loaded we need to pre-process the data , So for that we perform data exploration to see univariate data visualization related to the chance of churn data from guests which is shown in figure 4.

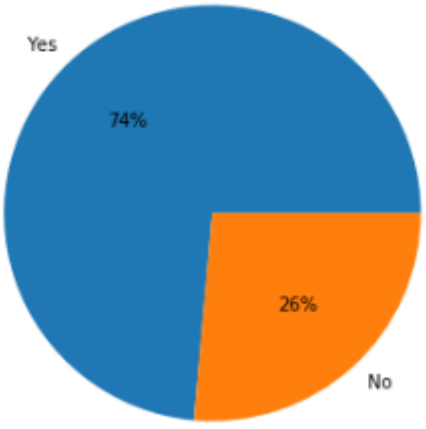

Figure 4. percentage of churn data from customers

Grounded on the results and analysis over, it can be concluded we know that the data distribution as a whole, the client doesn't churn, with details on Churn as much as 26 and No Churn as much as 74.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12744

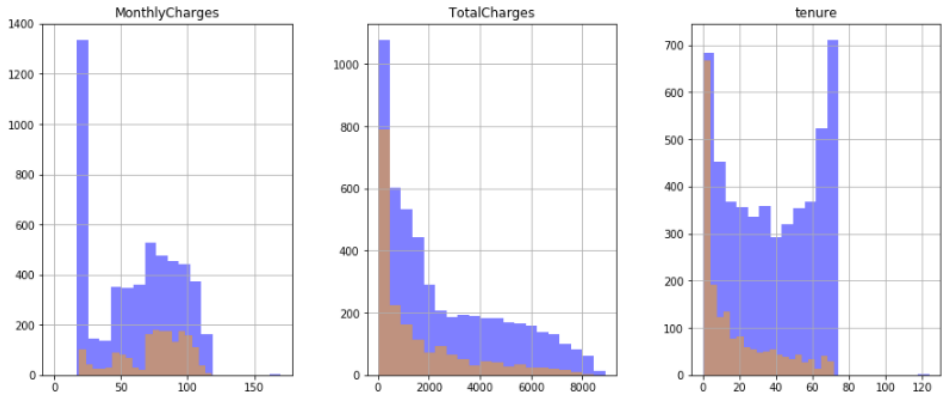

Figure 5. bivariate plot for numeric variable predictor

Based at the outcomes and evaluation above, it could be concluded: we will see that for MonthlyCharges there's a bent that the smaller the fee of the month-to-month expenses charged, the smaller the tendency to do Churn. For TotalCharges there doesn`t appear to be any inclination toward Churn customers. For tenure, there's a bent that the longer the client subscribes, the much less in all likelihood it's miles to churn.

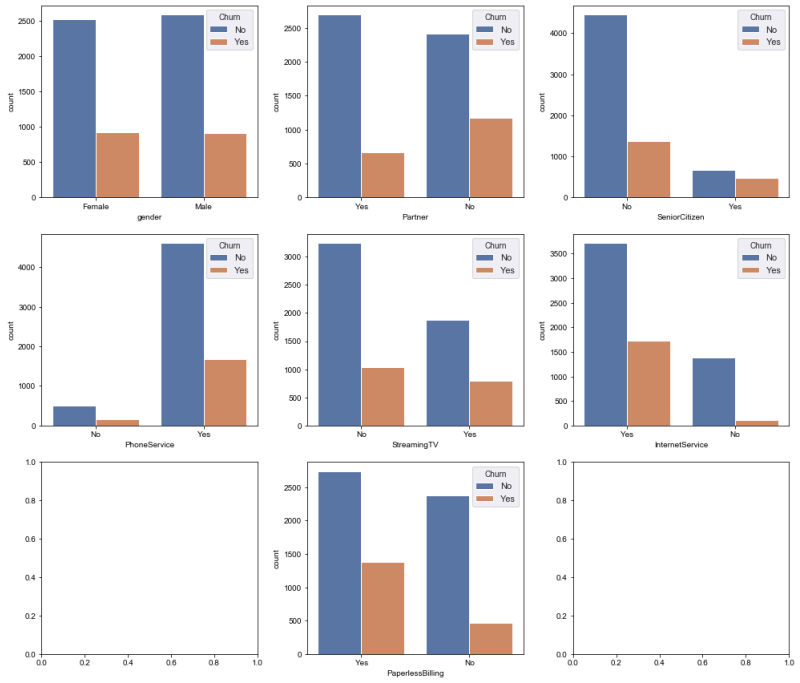

Figure 6. bivariate plot for categorical variable predictor

We recognise that there may be no vast distinction for humans doing churn in phrases of gender and smartphone service (Phone Service). However, there may be a bent that those who churn are those who do now no longer have a partner (partner: No), humans whose repute is a senior citizen (Senior Citizen: Yes), humans who've streaming TV services (StreamingTV: Yes) , humans who've Internet service (internetService: Yes) and those who've paperless bills (PaperlessBilling: Yes).

Pre-Processing Data

Before Building a fashions we will carry out pre-processing, Remove the needless columns customerID & UpdatedAt and Convert all of the non-numeric columns to numerical facts kinds than we will cut up the dataset 70% from schooling facts and ultimate 30% for checking out dataset.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12744

Building a Model

After Preprocessing we can build three Machine learning model : Logistic Regression, Random Forest and Gradient Boost Algorithm.

Logistic Regression

First we can take Logistic Regression and start building on training data and once the model gets trained we can test the performance of the model on test dataset and the result we have get are shown in figure 7.

```
# Performance Data Testing - Displays Metrics

# Predict
y_test_pred = log_model.predict(x_test)

# Print classification report
print(classification_report(y_test, y_test_pred))
```

```
              precision    recall  f1-score   support

           0       0.83      0.90      0.87      1539
           1       0.64      0.49      0.56       546

    accuracy                           0.80      2085
   macro avg       0.74      0.70      0.71      2085
weighted avg       0.78      0.80      0.79      2085
```

Figure 7. Prediction Result of Logistic Regression

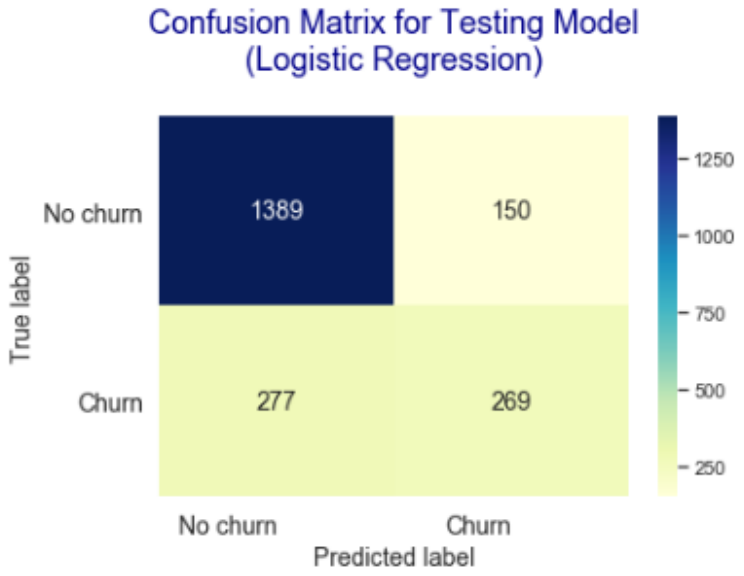

Figure 8. Confusion Matrix of Logistic Regression

Random Forest Classifier

Than we can take Random Forest Algorithm and start building on training data and once the model gets trained we can test the performance of the model on test dataset and the result we have get are shown in figure 9.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12744

```
# Performance Data Testing - Displays Metrics

# Predict
y_test_pred = rdf_model.predict(x_test)

# Print classification report
print(classification_report(y_test, y_test_pred))
```

```
              precision    recall  f1-score   support

           0       0.81      0.89      0.85      1539
           1       0.58      0.42      0.49       546

    accuracy                           0.77      2085
   macro avg       0.70      0.66      0.67      2085
weighted avg       0.75      0.77      0.76      2085
```

Figure 9. Prediction Result of Random Forest Algorithm

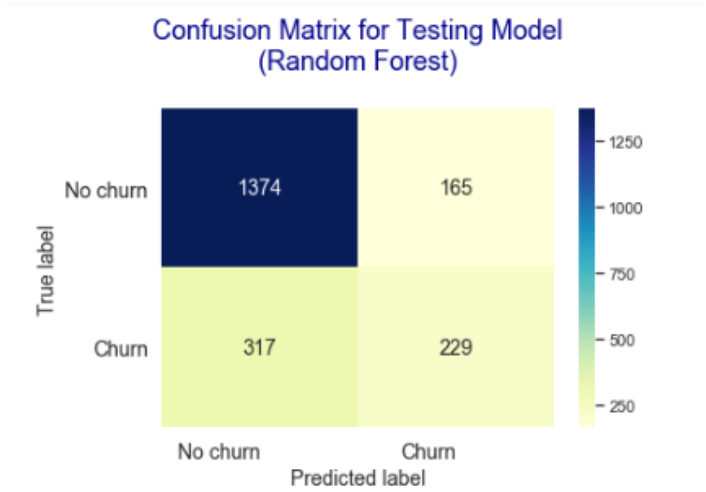

Figure 10. Confusion Matrix of Random Forest Algorithm

Gradient Boosting Classifier

Than we can take Gradient Boosting Algorithm and start building on training data and once the model gets trained we can test the performance of the model on test dataset and the result we have get are shown in figure 11.

```
# Predict
y_test_pred = gbt_model.predict(x_test)

# Print classification report
print(classification_report(y_test, y_test_pred))
```

```
              precision    recall  f1-score   support

           0       0.83      0.91      0.87      1539
           1       0.64      0.48      0.55       546

    accuracy                           0.79      2085
   macro avg       0.74      0.69      0.71      2085
weighted avg       0.78      0.79      0.78      2085
```

Figure 11. Prediction Result of Gradient Boosting Algorithm

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12744

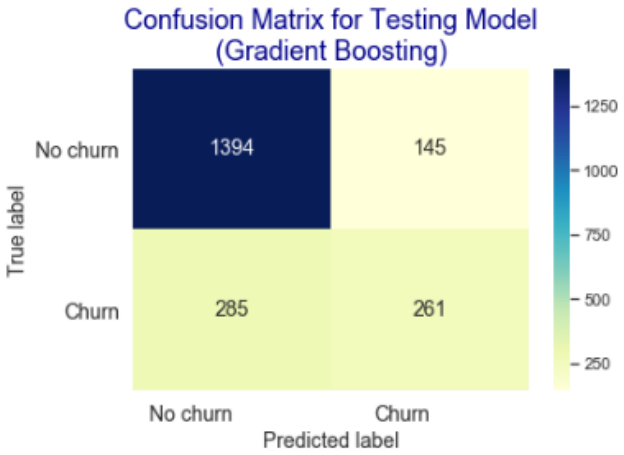

Figure 12. Confusion Matrix of Gradient Boosting Algorithm

Based on the Accuracy we have got from prediction result performance measures of the models we can compare the performance of the models and the conclusion of the performance are shown in figure 13.

Table 1. Accuracy Comparison

Algorithm	Accuracy
Random Forest	77%
Gradient Boosting	79%
Logistic Regression	80%

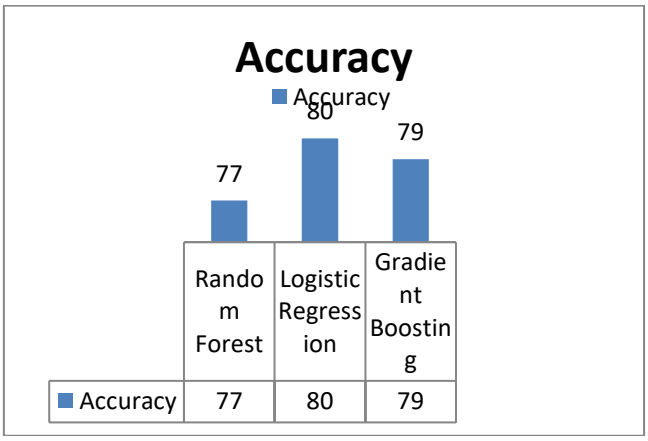

Figure 13. Performance on the models

VI. CONCLUSION

Analysis of statistics that's uprooted from telecom groups can facilitates to locate the motives of consumer churn and also makes use of the statistics to preserve the consumer. So prognosticating churn is usually important for telecom institutions to preserve their consumer. In order to preserve being guests, Telecom vendors want to understand the motives of churn, which may be found out thru the information uprooted from Telecom statistics. In this paper, we teach device literacy fashions that's Logistic Regression, Random Forest and Gradient Boosting, we will say that Logistic Regression is carry out higher as compared to Random Forest and Gradient Boosting because it provides better delicacy.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023
DOI: 10.17148/IJARCCE.2023.12744

REFERENCES

[1] V.Geetha, A.Nandhini, T. Nandhini, R.Sushmitha, " Customer Churn Prediction In Telecommunication Industry Using Random Forest Classifier" in 2021 IEEE.
[2] Mykola Malyar, Maryana Sharkadi, "Churn Prediction Estimation Based on Machine Learning Methods" in IEEE 2020.
[3] Dandi T. Barus , P. H. Gunawan, "Parallel Programming of Churn Prediction Using Gaussian Naive Bayes" in 2020 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.