World Journal of Advanced Engineering Technology and Sciences

eISSN: 2582-8266

Cross Ref DOI: 10.30574/wjaets

Journal homepage: https://wjaets.com/

(Research Article)

Check for updates

Telecom Churn Prediction using Machine Learning

Krishnan R [1], CV Krishnaveni [2, *] and AV Krishna Prasad [3]

[1] *Data Science and Engineering. Birla Institute of Technology and Science, Pilani (UCG, ACU, AIU Affiliated), India.*
[2] *Lecturer in Computer Science, SKR &SKR GCW, Kadapa, Andhra Pradesh, India.*
[3] *Maturi Venkata Subba Rao Engg. College, Hyderabad, Telangana, India*

World Journal of Advanced Engineering Technology and Sciences, 2022, 07(02), 087–096

Publication history: Received on 14 October 2022; revised on 22 November 2022; accepted on 25 November 2022

Article DOI: https://doi.org/10.30574/wjaets.2022.7.2.0130

Abstract

In every industry, customers are crucial. Customer churn can have a variety of effects and have a negative influence on sales. Analysis and forecasting of customer turnover must be a key component of any business. We will analyze and forecast customer turnover in the telecom industry in our study. The study of consumer behavior is crucial for the telecommunications sector in order to identify those customers who are most likely to cancel their subscriptions. Because there is so much data available and the market is becoming more competitive, businesses are spending more time trying to keep their present consumers than they are trying to win over new ones. The mobile telecommunications market recently transitioned from being one that was expanding quickly to one that was saturated. The goal of telecom companies is to refocus their attention away from attracting new, huge customers and toward retaining existing ones. Knowing which clients are likely to switch to a competitor in the future is important for this reason.

Using machine learning techniques such as Decision Tree, Logistic Regression, Random Forest, Gradient Boosted Machine Tree, and Extreme Gradient Boosting, the model is proposed for churn analysis and prediction for telecommunication firms. The performance of various models is also compared. On the basis of the supplied dataset, comparisons are made on the algorithm's effectiveness.

Keywords: Machine Learning; Variance Reduction; Prediction; Classification; Telecom; Churn; Logistic Regression; Bayesian Models; Random Forest; Gradient Boosted Machine Tree; Decision Tree

1. Introduction

In many countries, the telecommunications industry has grown to be important. The degree of competition increased as a result of both technological advancement and an increase in operators. Businesses are putting a lot of effort into surviving in this cutthroat market by utilizing various techniques. A vast amount of data is available as a result of the rapid expansion of the data transmission network and improvements in information technology. Customer churn is an important factor in the survival and growth of the telecom business because customers are the main source of profit. Here, businesses need to flourish, to survive in the cutthroat market by putting new plans and procedures in place for expanding their consumer base globally. The marketing team of the organization uses a variety of tactics to draw in new clients, encourage current clients to upgrade to new services offered by the same business, and ultimately maintain a clientele for an extended period of time.

Due to intense market competition, customers who are inclined to leave the telecom industry are a significant problem. On the other hand, if it is done early on, evaluating the consumers who are most likely to quit the business could offer a sizable extra revenue source. Numerous studies have demonstrated the effectiveness of machine learning algorithms in

[*] Corresponding author: CV Krishnaveni
Computer Science, SKR &SKR GCW, Kadapa, Andhra Pradesh, India.

predicting churning and non-churning events by learning from historical corporate data. All consumer data collected over time is included in the data used in this analysis.

This paper primarily focuses on various machine learning methods and algorithms for predicting churn in the telecom industries that are model-based and regression-based.

1.1. Understanding and Defining Churn

Customers in the telecom sector have access to a variety of service providers and can actively switch from one operator to another. The telecoms business has an average annual churn rate of **15 to 25** percent in this fiercely competitive market. Customer retention has now surpassed customer acquisition in importance due to the fact that it is **5–10 times** more expensive to gain new customers than to keep existing ones. For many business owners, retaining highly profitable consumers is their top priority.

1.1.1. Prepaid and Postpaid

In the telecom sector, there are two primary payment methods: postpaid (where consumers pay a monthly or annual fee after using the services) and prepaid (where users pay/recharge with a set amount in advance and then use the services).

In the postpaid model, consumers who desire to migrate to another operator typically notify the current operator to cancel the services, which might be identified as a case of churn.

In the prepaid model, users who desire to migrate to another network can abruptly stop using the services, making it difficult to determine if a client has genuinely left or is just temporarily not using services.

1.1.2. Build Predictive Models

Telecom businesses must identify the consumers who are most likely to migrate in order to reduce customer churn. This paper examines customer-level data from a top telecom company, builds predictive models to find customers who are likely to leave, and identifies the key churn factors. Therefore, for prepaid clients, churn prediction is typically more important and complex. This paper is focused on the telecom industry in India.

1.2. Type of Churns

1.2.1. Revenue-based Churn Customers

Customers who, for a certain amount of time, have not used any revenue generating services, such as mobile internet, outgoing calls, SMS, etc. The use of aggregate metrics is another option. For example, “A customers who have generated less than INR 3 per month in total/average/median revenue.” This definition’s critical problem is that some customers only get calls or SMS from their wage-earning equivalents. Customers in this case only consume the services, not create revenue.

1.2.2. Usage-based Churn

Customers that have not made any calls or used the internet in a while, either incoming or outgoing.

1.2.3. High-value Churn

~~About 80~~This paper builds its definition of high-value consumers on specific criteria and limits its attrition prediction to high-value clients. This definition may have the drawback that it may be too late to take corrective measures to retain the client once they have stopped utilizing the services for a period. Predicting churn may be ineffective if it is predicated on a “two-month zero usage” timeframe because by then the client will have already transferred to another operator.

1.3. Motivation and Goals

The objective of this paper is to provide a technique for telecom industry churn prediction. The expected outcome is an algorithm that pinpoints the customers who are most likely to switch operators. Despite the fact that numerous methods have been tested in recent years. The existing studies and methodology for predicting churn in the telecommunications industry still have a lot of scope for improvement.

2. Related Work

The paper [1], describes the steps involved in creating a decision support system using data mining. Qureshi team in [2] Lazarov team in [3] presented widely used data mining techniques for churn prediction. [4] suggests a new set of features to increase the recognition rates of potential churners.

[5] focuses on data mining strategies for reducing error ratio and customer churn. [6] analyses the many customer data classifications accessible in open datasets, as well as the performance measures and predictive models used in the domain in churn prediction in the telecom business. [7] established a customer classification and misclassification cost-based research model for customer churn. Additionally, a telecom company's customer behavior data was analyzed using this model. Furthermore, information on churn prediction can be found in the studies included in the [8],[9],[10],[11].

In [12], Kiran et al presented a state-of-art review of various methods and researches involve in churn prediction, have done assessments on frequently used data mining procedures to categorize customer churn patterns in telecom industry. The contemporary literature in the expanse of predictive data mining techniques in customer churn comportment is reviewed and categorized in terms of method used and an argument on the future research directions is presented. One of the latest papers, [13] worked on classification models using techniques like Logistic Regression, Random Forest, and Lazy Learning for predicting customer churn. Another latest work on Churn prediction [14] explained how to recommend bank customers using AI and ANN. They supported prior research that highlights the possibility of client loyalty.

3. Proposed Work

The proposed method entails a thorough examination and analysis of telecom datasets. The likelihood of that proportion churning is exceedingly difficult to estimate. Our solution makes it easier to understand why customers want to churn, and it will quickly display the data as bar plots and pie charts. The telecommunications business will benefit from the study of foreseeing who is going to depart the network and identify who will do so. The effectiveness of prediction results is measured using the techniques Logistic Regression, Random Forest, Gradient Boosted Machine Tree, Extreme Gradient Boosting, and Decision Tree. The proposed approach explains the system's work flow and the procedures involved.

3.1. Logistic Regression

One of the most crucial statistical methods used in data mining to conduct data analysis is logistic regression. Logistic regression is a generalized class notion of linear regression. To determine the likelihood of a target variable, a supervised learning classification algorithm is needed. LR belongs to a class of regression analysis techniques that are generally employed to identify and quantify correlations among dataset features. Regression analysis should be performed using the correct model when the dependent variable is binary. A predictive analysis called logistic regression is used to explain the relationship between a group of independent binary variables and a dependent binary variable. For customer churn, logistic regression has been used to calculate the likelihood of churn as a function of the traits or characteristics of the customers. Finding the likelihood of client churn also uses logistic regression. It is based on an approach to studying how variables affect other variables that is mathematically oriented.

3.1.1. Decision Tree

A supervised learning method called a Decision Tree can be used to solve classification and regression problems, but it is typically preferred for classification. The given dataset's features are used to execute the test or make the decisions. It is a graphical depiction for obtaining all potential answers to an issue based on predetermined parameters. The CART algorithm, which stands for Classification and Regression Tree algorithm, is used to construct a tree.

3.1.2. Random Forest

To forecast if a consumer will terminate his subscription, Random Forest is utilized. Decision trees are used by Random Forest to classify whether a consumer would cancel their subscription. Many different decision trees make up the random forest. A decision tree identifies a particular class. The classifier for a specific client will be the class with the most votes. The data that decision trees are educated on can affect how they behave. The use of bagging prevents this. Taking a random sample from the dataset to train the decision trees is a technique known as bagging.

3.1.3. XGBoost

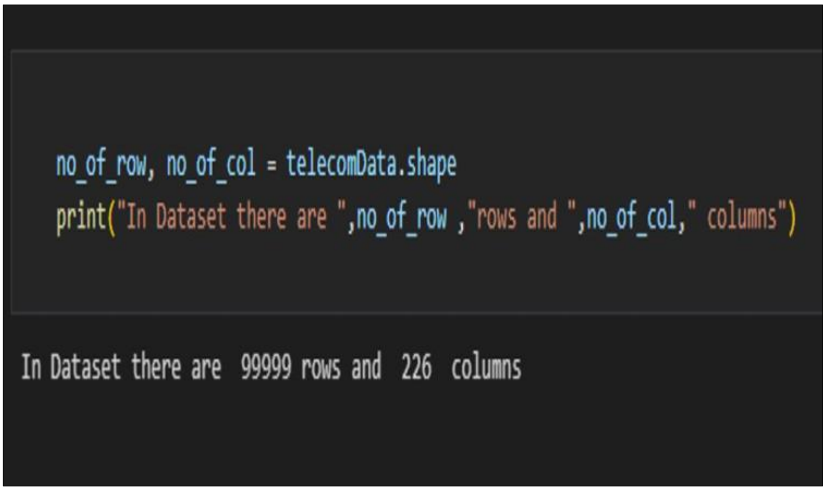
Extreme Gradient Boosting is referred to as XGBoost. The main reason for adopting XGBoost is because of how quickly it executes and how well the model's function. XGBoost employs ensemble earning techniques, which combine a number of different algorithms to obtain results from a single model. XGBoost provides optimal memory consumption while supporting distributed and parallel processing.

3.2. Data Preprocessing

The Telecom Churn dataset contains 226 columns Number of rows in the csv file = 99,999 It contains a lot of missing values data set available for this research is too large and hence needed to be sampled. In order to prevent the loss of valuable information and keep the final results more accurate. train and test sets are built by randomly splitting in the ratio of 70:30.

3.2.1. Elimination of Unique Values

Check for unique values and remove any columns that have them because they are not needed for analysis



```
no_of_row, no_of_col = telecomData.shape
print("In Dataset there are ",no_of_row ,"rows and ",no_of_col," columns")
```

```
In Dataset there are  99999 rows and  226  columns
```

Figure 1 Telecom Churn Dataset size

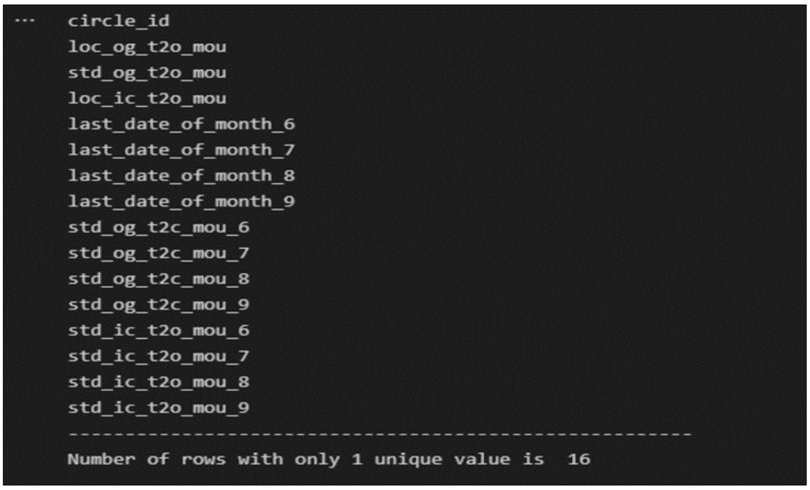
3.2.2. Handling of Missing Values

In the missing value columns, there are three types: Boolean Columns: Must be filled by 1 or 0. Columns with Boolean values are night pack, fbuser Date Columns: Must be filled by dates. Columns with dates are dateoflastrecharge Numerical Columns: Must be filled by numbers. Remaining all missing value columns

Further, derivation of features is one of the most crucial steps in data preprocessing since good features may frequently distinguish between good and bad models. Utilize your knowledge of business to derive characteristics that are thought to be key churn indicators.

3.2.3. Filtration of high-value customers

Churn prediction is only performed for high-value clients, as was already mentioned. As an example of a high-value client, consider those who recharged with a quantity greater than or equal to X, where X is the 70th percentile of the typical recharge amount during the first two months, which is the favorable phase. About 29.9k rows were obtained after excluding the high-value clients. Mark churners and take away churn phase characteristics now, based on the fourth month, tag the churned consumers as follows: Those who have neither placed or received any calls AND have not even once accessed mobile internet during the churn phase. The attributes used to tag churners are: totalicmou9 totalogmou9 vol2gmb9 vol3gmb9 after tagging churners, remove all the attributes corresponding to the churn phase.



```
...   circle_id
      loc_og_t2o_mou
      std_og_t2o_mou
      loc_ic_t2o_mou
      last_date_of_month_6
      last_date_of_month_7
      last_date_of_month_8
      last_date_of_month_9
      std_og_t2c_mou_6
      std_og_t2c_mou_7
      std_og_t2c_mou_8
      std_og_t2c_mou_9
      std_ic_t2o_mou_6
      std_ic_t2o_mou_7
      std_ic_t2o_mou_8
      std_ic_t2o_mou_9
      ----------------------------------------------------
      Number of rows with only 1 unique value is  16
```

Figure 2 Unique Values in Churn

3.2.4. Churn data Exploratory Analysis

The target label that has to be predicted must be determined before the analysis of the columns can continue. Predicting client attrition for the ninth (September) month is the objective given here. Therefore, a user is considered to have been churned if they don't make any calls or consume any data in the ninth month. Let's partition the columns and add churn columns to better execute exploratory data analysis. For analysis, it was determined which pair plots could distinguish between churned and non-churned customers:

Understanding-1

From the above figure3, it is understood that from the aforementioned plots, there is no clear way of distinguishing. There is a significant amount of overlap.

3.2.5. Observation of Features

According to the histograms from the pair plots, the revenue generated by the churned clients in the eighth month is typically very low. The distribution plot and box plot both show the same thing. Box plots show that the range of values for the eighth month may be distinguished. However, this is not true of the other months. A significant amount of the values for 6,7 months overlaps. Understanding-2: From the above figure4, it is understood that Features of the dataset seems much more distinguishable. Especially those that contain local calls from the eighth month.

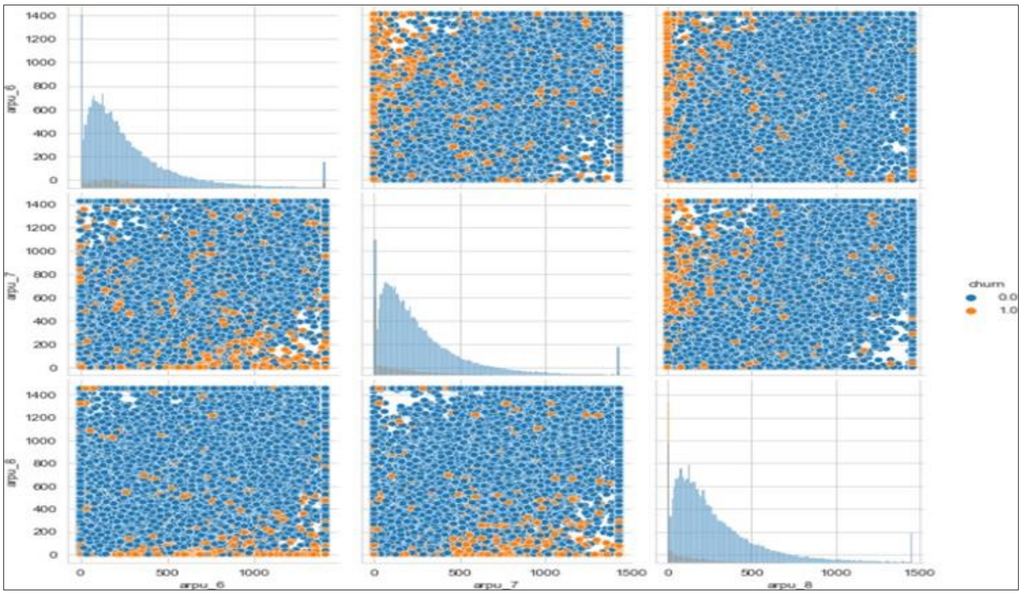


Figure 3 Data Exploration of Churn

3.3. Evaluation Metrics

Performance of traditional classification algorithms is evaluated by the metric accuracy which is defined as the percentage of examples that are correctly classified. This is not suitable when dealing with imbalanced data sets as the minority class has a smaller number of samples. In fact, misclassifying all minority samples and correctly classifying majority class samples gives a very good accuracy. Performance of a classifier is calculated based on the confusion matrix.

Table 1 Confusion Matrix

	Actual Positives	**Actual Negatives**
Predicted Positives	True Positives (TP)	False Positives (FP)
Predicted Negatives	False Negatives (FN)	True Negatives (TN)

Various measures used for describing the performance of the classifiers are listed below:

3.3.1. Sensitivity

Sensitivity is the percentage of Positives Correctly Classified. It denotes the accuracy of the positive class. Recall and True Positive Rate (TP Rate), TPR are other names of Sensitivity.

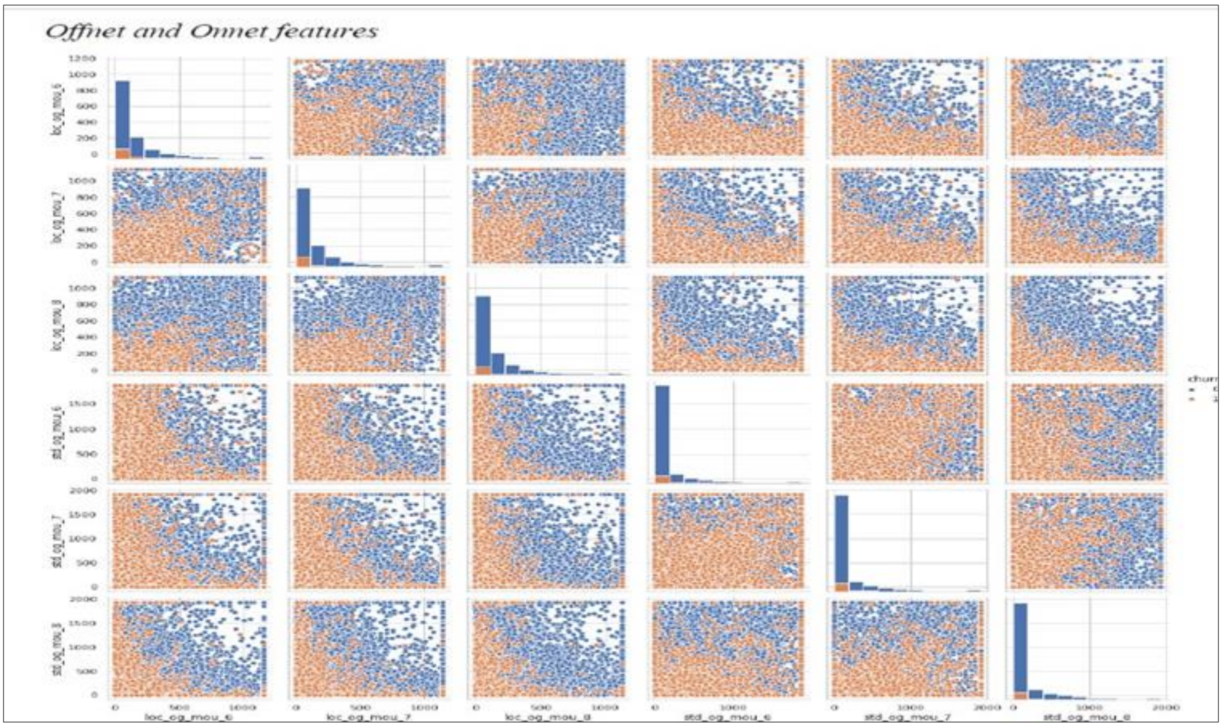


Figure 4 Features of Churn

$$\frac{Sensitivity\ =\ TP\ Rate\ =}{Recall\ =}\frac{TP}{TP\ +FN}$$

3.3.2. Specificity

Sensitivity is the percentage of Positives Correctly Classified. It denotes the accuracy of the negative class True Negative Rate (TN Rate), TNR are other names of Specificity.

$$Specificiy\ =\ TN\ Rate\ =\frac{TN}{TN\ +\ FP}$$

3.3.3. False Positive Rate

False Positive Rate is the percentage of negatives wrongly classified.

$$FP\ Rate\ = \frac{FP}{TP\ +\ FN}$$

3.3.4. False Negative Rate

False Negative Rate is the percentage of positives wrongly classified.

$$FN\ Rate\ \frac{FN}{TN\ +\ FP}$$

Accuracy: The percentage of correctly classified instances.

$$\text{Accuracy}\ = \frac{(TP\ +\ TN\ )}{(TP\ +\ FN\ +\ TN\ +\ FP\ )}$$

Error Rate: The Percentage of incorrectly classified instances.

$$Error\ Rate\ = \frac{(FP\ +\ FN\ )}{(TP\ +\ FN\ +\ TN\ +\ FP\ )}$$

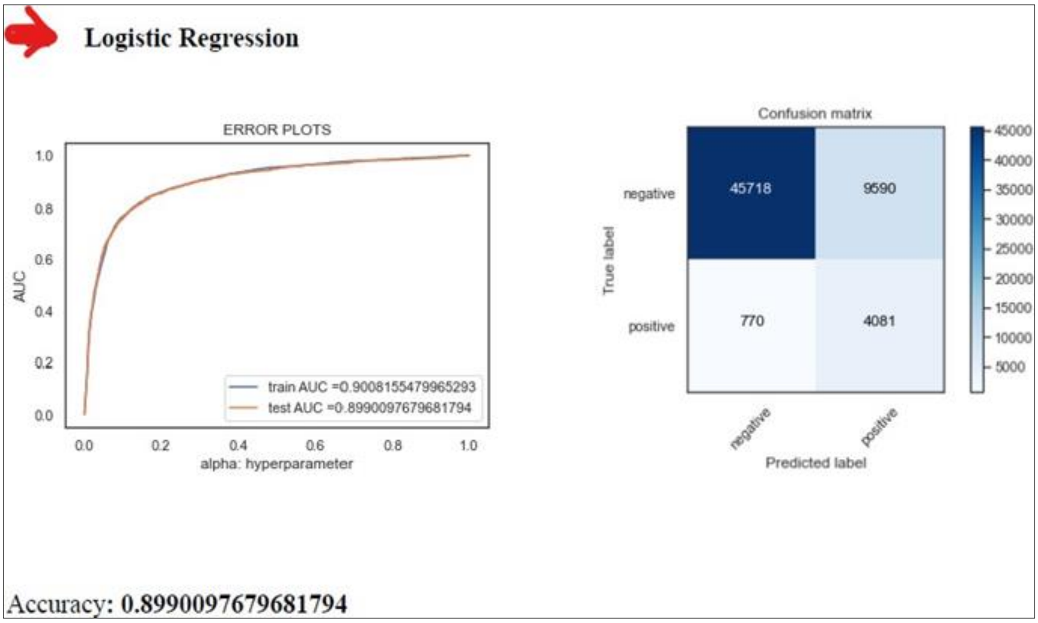
GMean: It is the geometric mean of Sensitivity and Specificity.

$$Gmean\ = \frac{\surd}{Sensitivity\ \times\ Specificity}$$

F-Measure: It is the harmonic mean of Precision and Recall.

$$F - Measure\ = \frac{2\ \times\ Precision\ \times\ Recall}{Precision\ +\ Recall}$$

In this paper, Accuracy measure is used to know the performance of the model.



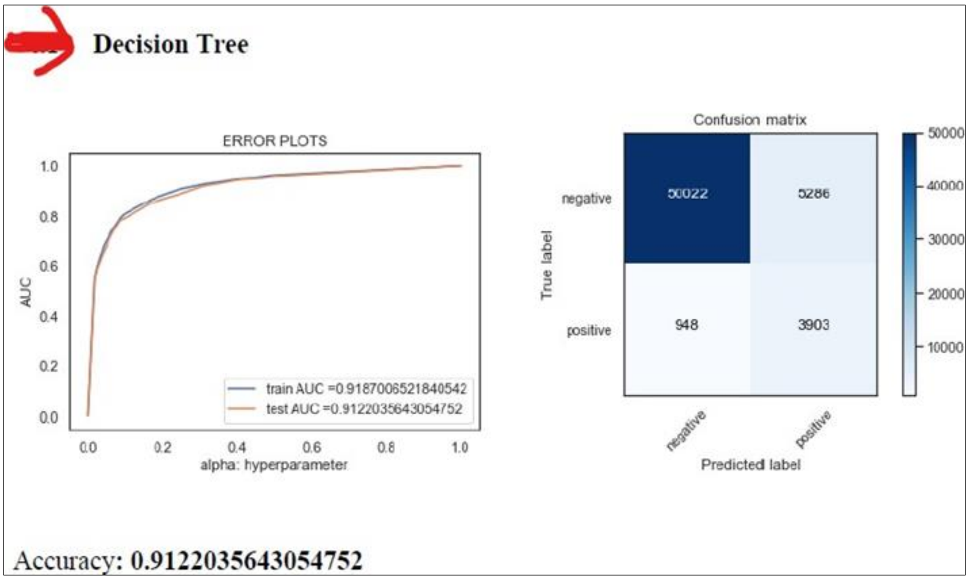
Logistic Regression

Accuracy: **0.8990097679681794**

Figure 5 Logistic Regression Results

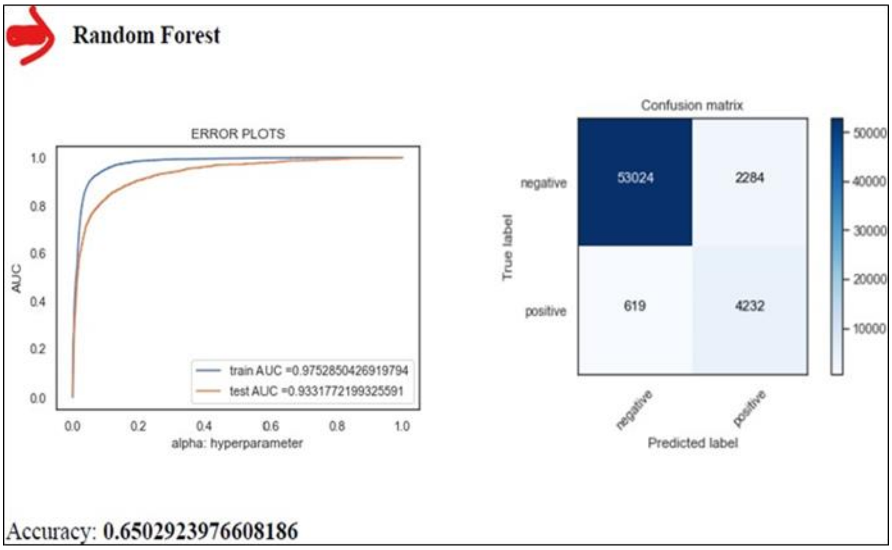
3.4. Model Building and Results

After thorough examination of all the important features of the dataset, the most prevalent finding from these is that the number of churned customers is large in the sixth and seventh months but declines for the eighth month. This is typical for the majority of churned consumers, but not all. There are significant links between the sixth and eighth months. Recharge amounts are easier to discern from the other features. The points overlap for the majority of the features. As a result, a distinct distinction was not possible. The majority of churned consumers have poor values in particular during the eighth month. The pair plots demonstrate that the majority of churned customers have values in either of the two columns that are close to 0. In the above figures 5,6, 7,8 results obtained from the models, confusion matrix and accuracy, and AUC curve are depicted. The table 2 shows the comparison of accuracy of machine learning models implemented in this paper and it is proved that XGBoost outperforms other models in prediction of Telecom Churn.



Accuracy: 0.9122035643054752

Figure 6 Decision Tree Results

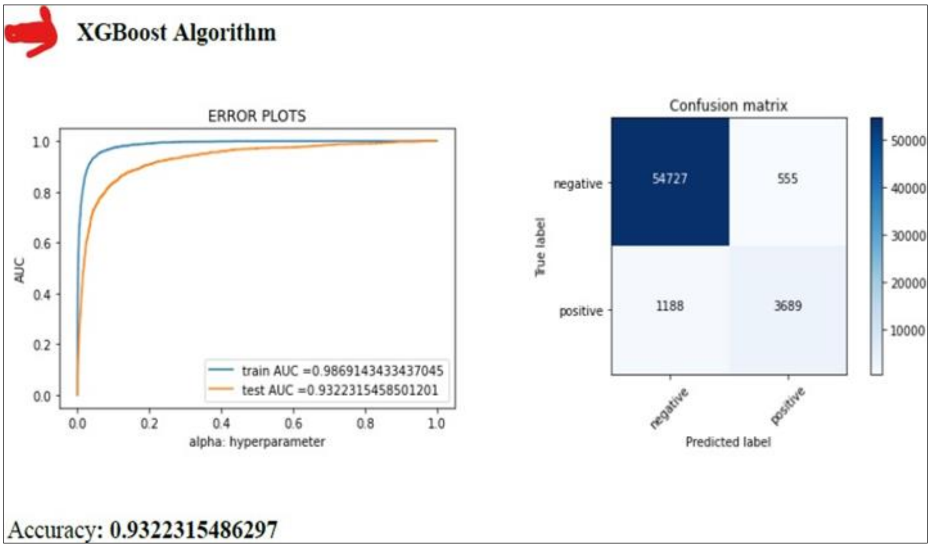


Accuracy: 0.6502923976608186

Figure 7 Random Forest Results

Table 2 Comparison of Results of Churn Prediction by various models

Model	**Accuracy**
Logistic Regression	0.89
Decision Tree	0.91
Random Forest	0.65
XGBoost	0.93


XGBoost Algorithm

Accuracy: **0.9322315486297**

Figure 8 XGBoost Results

4. Conclusion

The issue of client turnover has gotten much worse as the telecommunications sector has continued to expand. In the telecommunications sector, customer retention is a crucial concern since it lowers customer churn by raising customer satisfaction. The classification of the network-leaving clients will be advantageous to the telecommunications industry. This problem can be solved using machine learning algorithms and predictive analytics. The study examined machine learning techniques and applied to use on a dataset. By modelling and testing, the assessment metrics for the Logistic Regression, Random Forest, Gradient Boosted Machine Tree, Extreme Gradient Boosting and Decision Tree models were improved. This work can be extended to various other datasets from the telecom department and implement methodologies to achieve better results in future work.

Compliance with ethical standards

Acknowledgments

Many thanks to everyone who took part in the study and made it possible for this research paper.

Disclosure of conflict of interest

We have declared that no conflict of interest.

References

[1] Jadhav, Rahul Pawar, Usharani. (2011). Churn Prediction in Telecommunication Using Data Mining Technology. International Journal of Advanced Computer Sciences and Applications. 2.10.14569/IJACSA.2011.020204.

[2] Qureshi, Saad Rehman, Ammar Qamar, Ali Kamal, Aatif Rehman, Ahsan. (2013). Telecommunication Subscribers' Churn Prediction Model Using Machine Learning. 8th International Conference on Digital Information Management, ICDIM 2013. 10.1109/ICDIM.2013.6693977.

[3] Lazarov, Vladislav and Marius Capota. "Churn Prediction." (2007).

[4] Kirui, Clement K. et al. "Predicting Customer Churn in Mobile Telephony Industry Using Probabilistic Classifiers in Data Mining." (2013)

[5] Balasubramanian, Madhana Veerapandian and M. Pradhiba Selvarani. "CHURN PREDICTION IN MOBILE TELECOM SYSTEM USING DATA MINING TECHNIQUES." (2014).

[6] Umayaparvathi, V. and K. Iyakutti. "A Survey on Customer Churn Prediction in Telecom Industry: Datasets, Methods and Metrics." (2016).

[7] Liu, Y. and Yongrui Zhuang. "Research Model of Churn Prediction Based on Customer Segmentation and Misclassification Cost in the Context of Big Data." Journal of Computational Chemistry 03 (2015): 87-93.

[8] Canale, Antonio and Nicola Lunardon. "CHURN PREDICTION IN TELECOMMUNICATIONS INDUSTRY. A STUDY BASED ON BAGGING CLASSIFIERS." (2014).

[9] Hashmi, Nabgha Butt, Naveed Anwer Iqbal, Dr.Muddesar. (2013). Customer Churn Prediction in Telecommunication A Decade Review and Classification. IJCSI. 10. 271-282.

[10] L. F. Khalid, A. Mohsin Abdulazeez, D. Q. Zeebaree, F. Y. H. Ahmed and D. A. Zebari, "Customer Churn Prediction in Telecommunications Industry Based on Data Mining," 2021 IEEE Symposium on Industrial Electronics Applications (ISIEA), 2021, pp. 1-6, doi: 10.1109/ISIEA51897.2021.9509988

[11] Babu, Pr. Sathesh et al. "A Review on Customer Churn Prediction in Telecommunication Using Data Mining Techniques." (2016).

[12] Kiran Dahiya Kanika Talwar." Customer Churn Prediction in Telecommunication Industries using Data Mining Techniques- A Review." (2015) ISSN: 2277 128X, Volume 5, Issue 4, pp 417-433.

[13] B, Senthilnayaki M, Swetha D, Nivedha. (2021). CUSTOMER CHURN PREDICTION. IARJSET. 8. 527-531. 10.17148/IARJSET.2021.8692.

[14] Y. Kavyarshitha, V. Sandhya and M. Deepika," Churn Prediction in Banking using ML with ANN," 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), 2022, pp. 1191-1198, doi: 10.1109/ICICCS53718.2022.9788456.