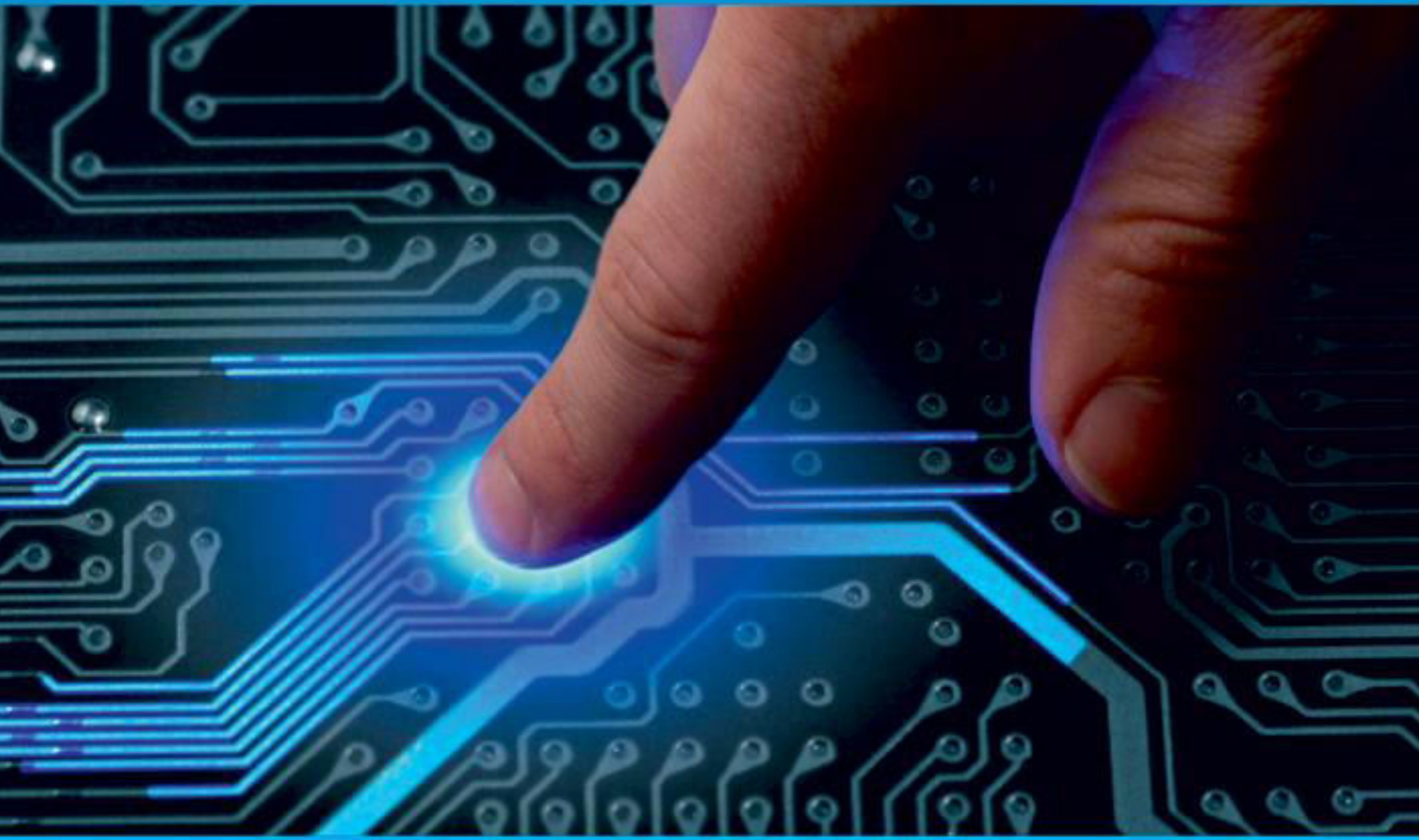
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 5, May 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005245|

A Novel Approach for Attrition Analysis in Telecom industry

Prajwal S P[1], Mayur S Sakhare[2], Manoj Badiger[3], Prabhakar Kulkarni[4], Prof.Radhika Patil[5]

B.E Student, Department of Computer Science and Engineering, Bapuji Institute of Engineering &Technology, Davangere Bangalore, Karnataka, India[1]

B.E Student, Department of Computer Science and Engineering, Bapuji Institute of Engineering &Technology, Davangere Bangalore, Karnataka, India[2]

B.E Student, Department of Computer Science and Engineering, Bapuji Institute of Engineering &Technology, Davangere Bangalore, Karnataka, India[3]

B.E Student, Department of Computer Science and Engineering, Bapuji Institute of Engineering &Technology, Davangere Bangalore, Karnataka, India[4]

Professor Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davangere, Karnataka, India[5]

ABSTRACT: Customers are the most valuable assets in any business since they are the primary source of profit. In today's competitive economy, businesses must rely on a variety of techniques in order to stay afloat. Increasing client retention, gaining new customers, and upselling current customers are the three basic ways to increase income. Upselling is more difficult than maintaining an existing client, yet the comparison of the two shows that customer retention costs far less than customer acquisition. Companies must reduce customer attrition in order to use the third approach. Customers who stop doing business with a firm are said to have "churned," which is the technical word for this phenomenon. When it comes to the number of consumers or clients who leave a firm, this is known as the churn rate. One of the most pressing issues for major corporations is customer attrition. Companies, particularly those in the telecommunications industry, are scrambling to find ways to anticipate clients who could cancel their service. As a result, it's critical to identify the root causes of client attrition in order to take the required steps to cut down on the attrition. During this project, we built a churn prediction model that may help businesses identify customers who are most likely to leave. There are a variety of machine learning approaches used to find the key reasons for customer churn and the best algorithm for making such predictions, including Logistic Regression, Modified Random Forest, Decision Trees, K-Nearest Neighbors, and Support Vector Machine algorithms. Customers' demographic information, total costs, and service type are all included in the dataset. Over 7000 customers' turnover data from 21 different Kaggle characteristics make up this report. Predicting customer attrition may also be done using the techniques discussed above in this study.

KEYWORDS: Logistic Regression, Modified Random Forest, Decision Trees, K-Nearest Neighbors, and Support Vector Machine.

I. INTRODUCTION

Telecom is one of the world's most important industries. There is now a much higher degree of competition because of the rapid expansion in technological and subscriber numbers. As a result of this fierce competition, a number of strategies have been devised by telecom companies to bring in money. It is critical for firms to reduce the rate of customer turnover in order to improve client retention. the moving of a consumer from one service provider to another service provider is referred to as "churn" (Ascarza, et al 2016). In service industries where competition is fierce, client turnover is a key concern (Ahmad, et al 2019). Many research have shown that machine learning systems are becoming more and more successful in predicting this scenario.

Customer churn prediction in the telecommunications sector is based on estimating the percentage of current customers who will stop doing business with a certain provider in the near future. and provide suggestions for preventing significant turnover. In today's competitive business climate, predicting churners before they leave has become more important. In light of the telecom industry's prominence, it was imperative to develop prediction techniques in addition to churn forecasting. Few studies have examined the importance of user retention in this sector. There is evidence to suggest that an increase in the overall value of a company's stock of only 1 percent might lead to a 5 percent increase in customer retention (Kisiogluand Topcu 2010). As Peppard (2000) claims, the rise of electronic commerce has increased the availability of information, and as a result, the internet channel has empowered the customers who are no longer stuck with the decisions of one company and has led to an exacerbation of the competition, while competitors are only one "click away" (Lejeune, 2001). The most efficient and effective ways to examine the behaviour of their clients and forecast their potential future failure should be available to enterprises confronting this challenge.

Using machine learning methods, we want to develop a model that can accurately forecast customer attrition in the prepaid mobile phone market sector. Data on customer turnover in telecoms is presented first in order to familiarise the reader with research's scope and relevance, and then we address our issue description and the subject of our investigation.

II. LITERATURE REVIEW

Rosa,2019 [1], "Artificial Neural Network for assessing and predicting customer attrition in the banking industry" suggested a new framework for assessing and predicting customer attrition in the banking industry that used Artificial Neural Network. The bank's Data Warehouse was used to get information about 1588 customers from January 2017 to December 2017. Only making neural networks was the point of the study. Other machine learning methods, like Decision Trees, Logistic Regression, and Support Vector Machines, were not taken into account. The goal of this project is to find a new way to find people who might leave so that marketing plans can be made to keep them.

The 2017 paper "A churn prediction model for prepaid customers in telecom using fuzzy classifiers" by Muhammad Azeem, Muhammad Usman, and A. C. M. describes a fuzzy-based churn prediction model that has been proposed and tested using real data from a telecom company in South Asia. Several popular classifiers, such as Neural Networks, Linear Regression, C4.5, Support Vector Machines, AdaBoost, Gradient Boosting, and Random Forest, have been compared to fuzzy classifiers to show that fuzzy classifiers are better at predicting the right set of churners.

J. Vijaya and E. Sivasankar,2019 [3] The paper "Customer churn prediction through particle swarm optimization-based feature selection model with simulated annealing" said that the method uses particle swarm optimization (PSO) and suggests three types of PSO for churn prediction: PSO with feature selection as its pre-processing mechanism, PSO with simulated annealing, and PSO with both feature selection and simulated annealing. We compared the proposed classifiers with decision tree, naive bayes, K-nearest neighbour, support vector machine, random forest, and three hybrid models to see how well they could predict and how well they worked. Experiments have shown that meta-heuristics work better and are easier to predict.

Matrin Fridrich, 2017 [4] The prediction model is made to find customers who are likely to leave, according to the article "Hyperparameter Optimization of Artificial Neural Network in Customer Churn Prediction." The proposed model makes it easier to predict when a customer will leave by looking at things like TP rate, FP rate, and accuracy. The analysis is done on a labelled e-commerce retail dataset with information about the 10,000 customers with the highest CLV. A genetic algorithm is used to look for good parameter settings in the proposed hyperparameter search space in order to find the best ANN (Artificial Neural Network) classification model. A conditional inference tree structure looks at the big picture of the given optimization to analyse a part of the hyperparameter search space that has already been looked at. This helps figure out what the most important parts of a good ANN classification model are.

Gordini and Veglio, 2015, p. 5, "Customers Who Leave and Marketing Strategies to Keep Them. A use of support vector machines in the B2B-commerce industry that is based on the AUC parameter-selection technique" Parameters like how recent, how often, how long, what product category, failure, money, age, profession, gender, request status, etc. were used to compare performance. It was found that the proposed method was better at making predictions than Linear Regression, Neural Networks, and SVM, especially when the data was noisy, uneven, or not linear. So, their results show that the data-driven way of predicting churn and making strategies for keeping customers is better than the shortcuts that are often used in the B2B e-commerce industry for management.

Yongbin Zhang, Ronghua Liang, Yeli Li,2011 [6] In "Behavior-Based Telecommunication Churn Prediction with Neural Network Approach," a behavior-based system for predicting when a telecom customer will leave is explained. The way the system works is also explained in this paper. Traditional methods of predicting customer churn use customer demographics, contractual data, customer service logs, call-details, complaint data, bills, and payments as inputs and churn as the target output. This system, on the other hand, uses a clustering algorithm and only customer service usage information to predict customer churn. It can solve problems that traditional methods can't, such as missing or unreliable data and the fact that some inputs are related to others. This study shows a new way to solve a problem that has been around for a long time.

Femina Bahari and SudheepElayidom, 2015, [7] The paper "An Efficient CRM-Data Mining Framework for the Prediction of Customer Behavior" says that the UCI dataset came from the direct marketing campaigns of a Portuguese bank. The model is used to predict how customers will behave so that better decisions can be made about how to keep good customers. Researchers looked at two classification models, Nave Bayes and Neural Networks, and found that Neural Network was better than Nave Bayes algorithm in terms of accuracy and specificity, but Nave Bayes algorithm was better in terms of sensitivity.

Lu et al [8], The article "The Use of Boosting Algorithm to Predict Customer Churn in the Telecommunication Industry" says that customers were put into two groups based on the weights given by the boosting algorithm. The used data set came from a company that deals with telecommunications. It had about 700 variables and a group of active mobile customers. Boosting works better than logistic regression, and it does a good job of separating churn data.

Chih-Fong Tsai,Yu-Hsin Lu, 2009 [9] "Customer Churn Prediction by Hybrid Neural Networks" says that churn management is a big job for businesses and that they need to be able to predict customer churn to keep their best customers. Neural networks have been shown in the literature to be useful for predicting churn. On the other hand, it has been shown that hybrid data mining techniques that combine two or more techniques work better than many single techniques for a wide variety of problems. This paper looks at two hybrid models made by combining two different ways to predict customer churn: back-propagation artificial neural networks and self-organizing maps. The hybrid models are made by putting together ANN and ANN and SOM and ANN.

Taiwo & Adeyemo, 2012 [10] "Churn prediction in a telecommunication company" says that the paper uses descriptive and predictive data mining techniques to find out how subscribers call and who is likely to leave a telecommunication company. During the descriptive stage, customers were put into groups based on how they used the service. As algorithms for grouping things together, K-Means and Expected Maximization were used (EM). In the predictive stage of Weka, the Decisions ump, M5P, and RepTree classifiers were used. EM does better than Kmean in the descriptive stage, but M5P does better than both Decisions ump and RepTree in the predictive stage.

Wang et al, 2018 [11],"Prediction of customer churn in search ads" says that the paper worked on a large-scale ensemble model to predict customer churn in search ads. The study's goal was to find out which customers were most likely to stop using the ads platform. The ensemble model of gradient boosting decision tree (GBDT) was used to predict whether or not a customer would leave in the near future based on how they interact with search ads. The data set came from the Bing Ads platform, and the results showed that the static and dynamic features work well together (AUC=0.8410).

He et al, 2014 [12],"Customer attrition analysis of commercial bank using Support Vector Machine" means that the paper did research on customer attrition analysis of commercial bank using Support Vector Machine. The 50000 customer records for the dataset came from the Chinese commercial bank data warehouse. 46,406 records were used to model after missing values and outliers were taken out. The SVM algorithm was used, but because the dataset wasn't balanced, the Random sampling method was added to make SVM better because it has a higher degree of recognition. When random sampling and the support vector machine algorithm were used together, the results showed that the predictive power went up a lot and churn in grate could be accurately predicted.

Based on the above survey, we have a good idea of how to predict customers leaving the telecom industry. Churning is the process of a customer switching from one company to another. It can be done with logistic regression, random forest, or support vector machine (SVM). When the above algorithms are used to train a data set, the trained model always stores the best prediction of churn. Realistic sequences with long-range structures can use it. The main goal of this model is to predict whether or not a customer will leave.

III. PROBLEM STATEMENT

Keeping customers from leaving is one of the most difficult things for businesses to do, especially those that offer subscription-based services. The loss of customers is called "customer churn" or "customer attrition." It can be caused by things like a change in preference, not having a good plan for dealing with customers, geographic change or other things. If businesses can accurately predict customer turnover, they can divide customers into groups based on how likely they are to leave and offer better services to keep them.

Proposed Solution

The data acquisition for the attrition analysis is done from the kaggle repository where the dataset consists information of over 7000 customers and 21 attributes like the customer name, how much they pay and what kind of service they get from the company are there in the dataset. we will use different machine learning models that show how the predictors relates to the response try to figure out what customers will do and we will also compare different machine learning models based on different performance evaluation metrics find out which is the most efficient machine learning model .

System Design

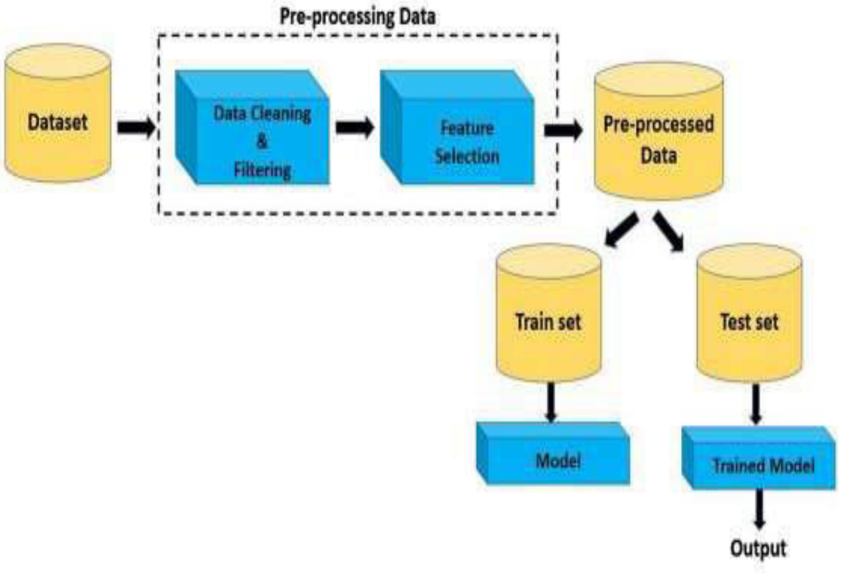


Fig 1: System Architecture Diagram

System architecture consists of the following steps:

- Collecting dataset appropriate for the project.
- Preprocessing the data for better results.
- Splitting them into training and testing data.
- Designing different machine learning and deep learning models for prediction.
- Predict price using different models and choose the model with the highest accuracy.

IV. METHODOLOGY

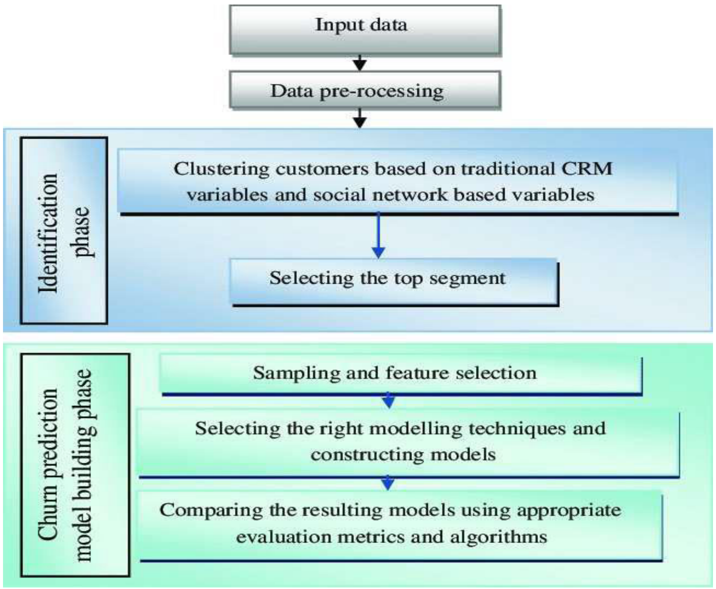

Input data
Data pre-rocessing
Identification phase
Clustering customers based on traditional CRM variables and social network based variables
Selecting the top segment
Churn prediction model building phase
Sampling and feature selection
Selecting the right modelling techniques and constructing models
Comparing the resulting models using appropriate evaluation metrics and algorithms

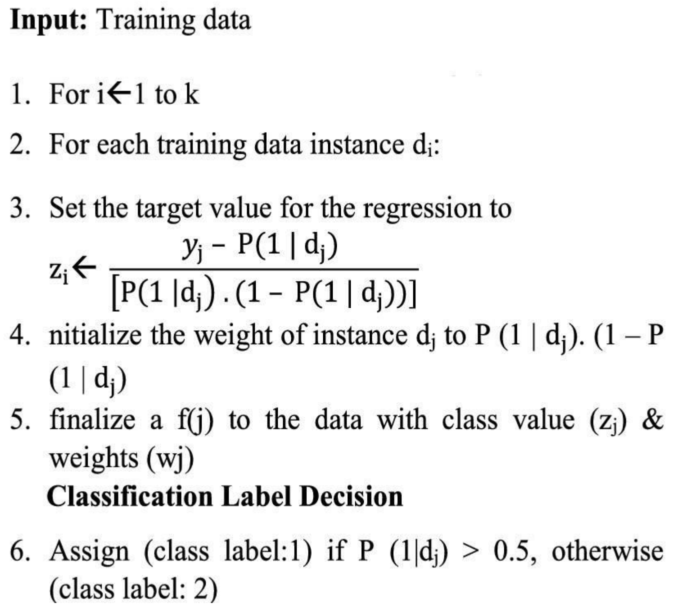
fig. 2. Methodology Diagram

The methodology proposed involves the following steps:

- Collect the required dataset and perform the pre-processing of collected data in order to remove null values, and normalize and scale data as required by the model.
- Once the preprocessing is done, the data is divided into training data and testing data.
- The training data is used to train the model in all aspects and testing data tests the trained model so that it gives the correct result.
- This methodology comprises of each model is trained and tested individually for the collected dataset based on which the model are trained and their accuracies are compared.

Pseudo code

Logistic regression



```
Input: Training data

1. For i ← 1 to k
2. For each training data instance $d_i$:
3. Set the target value for the regression to
   $z_i \leftarrow \frac{y_j - P(1 \mid d_j)}{[P(1 \mid d_j) . (1 - P(1 \mid d_j))]}$
4. nitialize the weight of instance $d_j$ to P (1 | $d_j$). (1 – P (1 | $d_j$)
5. finalize a f(j) to the data with class value ($z_j$) & weights (wj)
   Classification Label Decision
6. Assign (class label:1) if P (1|$d_j$) > 0.5, otherwise (class label: 2)
```

K-Nearest Neighbour Classifier

```
1. Load the training and test data
2. Choose the value of K
3. For each point in test data:
   - find the Euclidean distance to all training data points
   - store the Euclidean distances in a list and sort it
   - choose the first k points
   - assign a class to the test point based on the majority of classes present in the chosen points
4. End
```

Decision Tree Classifier

1.It begins with the original set S as the root node.
2.On each iteration of the algorithm, it iterates through the very unused attribute of the set S and calculates Entropy(H) and Information gain(IG) of this attribute.
3.It then selects the attribute which has the smallest Entropy or Largest Information gain.
4.The set S is then split by the selected attribute to produce a subset of the data.
5.The algorithm continues to recur on each subset, considering only attributes never selected before.

Random Forest Classifier

1: In Random Forest n number of random records are taken from the data set having k number of records.
2: Individual decision trees are constructed for each sample.
3: Each decision tree will generate an output.
4: Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

AdaBoost Classifier

1.Assign Equal Weights to all the observations.

2.Classify random samples using stumps.

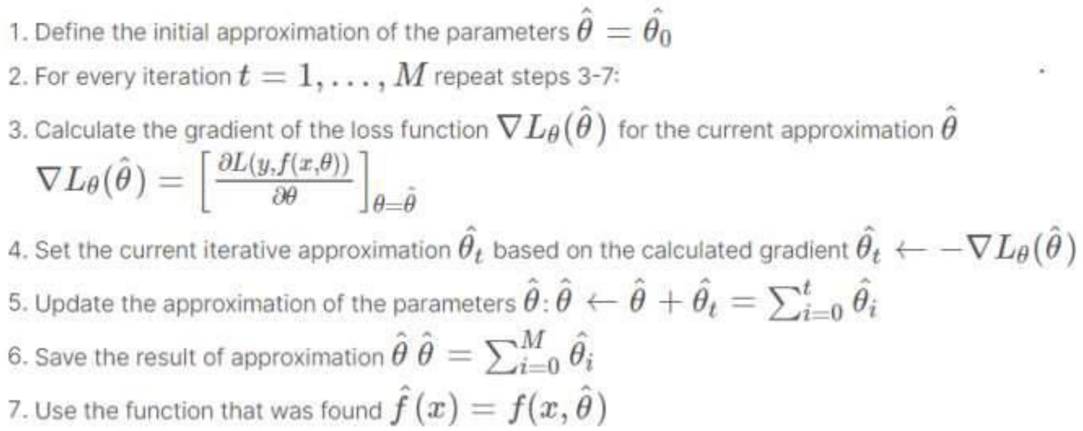
3. Calculate Total Error.

4.Calculate Performance of the Stump.

5. Update Weights.

1) **6** .Update weights in iteration.
7 .Final Predictions.

Gradient Boost Classifier



1. Define the initial approximation of the parameters $\hat{\theta} = \hat{\theta}_0$

2. For every iteration $t = 1, \dots, M$ repeat steps 3-7:

3. Calculate the gradient of the loss function $\nabla L_\theta(\hat{\theta})$ for the current approximation $\hat{\theta}$

$$\nabla L_\theta(\hat{\theta}) = \left[\frac{\partial L(y, f(x,\theta))}{\partial \theta}\right]_{\theta=\hat{\theta}}$$

4. Set the current iterative approximation $\hat{\theta}_t$ based on the calculated gradient $\hat{\theta}_t \leftarrow -\nabla L_\theta(\hat{\theta})$

5. Update the approximation of the parameters $\hat{\theta}$: $\hat{\theta} \leftarrow \hat{\theta} + \hat{\theta}_t = \sum_{i=0}^{t} \hat{\theta}_i$

6. Save the result of approximation $\hat{\theta}$ $\hat{\theta} = \sum_{i=0}^{M} \hat{\theta}_i$

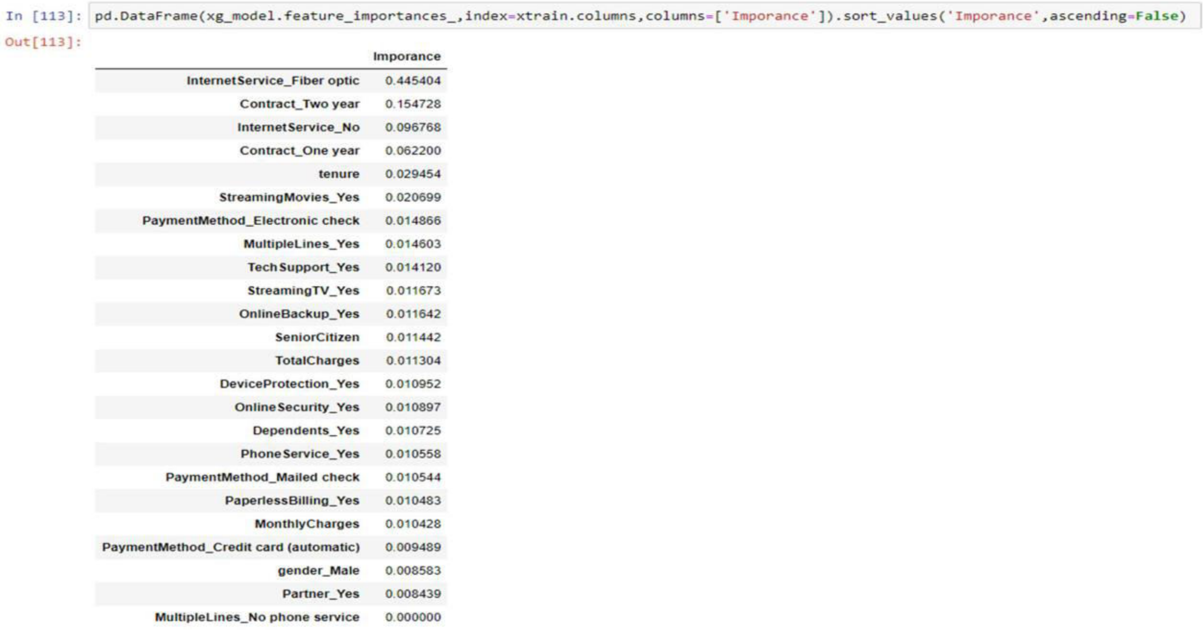
7. Use the function that was found $\hat{f}(x) = f(x, \hat{\theta})$

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005245|

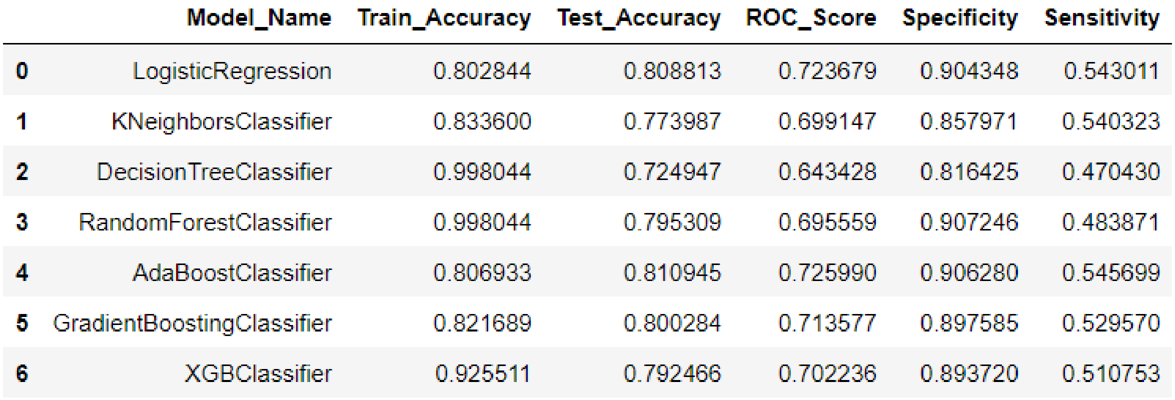
V. RESULTS AND DISCUSSION



```
In [113]: pd.DataFrame(xg_model.feature_importances_,index=xtrain.columns,columns=['Imporance']).sort_values('Imporance',ascending=False)
Out[113]:
```

	Imporance
InternetService_Fiber optic	0.445404
Contract_Two year	0.154728
InternetService_No	0.096768
Contract_One year	0.062200
tenure	0.029454
StreamingMovies_Yes	0.020699
PaymentMethod_Electronic check	0.014866
MultipleLines_Yes	0.014603
TechSupport_Yes	0.014120
StreamingTV_Yes	0.011673
OnlineBackup_Yes	0.011642
SeniorCitizen	0.011442
TotalCharges	0.011304
DeviceProtection_Yes	0.010952
OnlineSecurity_Yes	0.010897
Dependents_Yes	0.010725
PhoneService_Yes	0.010558
PaymentMethod_Mailed check	0.010544
PaperlessBilling_Yes	0.010483
MonthlyCharges	0.010428
PaymentMethod_Credit card (automatic)	0.009489
gender_Male	0.008583
Partner_Yes	0.008439
MultipleLines_No phone service	0.000000

Fig: Feature Importance



	Model_Name	Train_Accuracy	Test_Accuracy	ROC_Score	Specificity	Sensitivity
0	LogisticRegression	0.802844	0.808813	0.723679	0.904348	0.543011
1	KNeighborsClassifier	0.833600	0.773987	0.699147	0.857971	0.540323
2	DecisionTreeClassifier	0.998044	0.724947	0.643428	0.816425	0.470430
3	RandomForestClassifier	0.998044	0.795309	0.695559	0.907246	0.483871
4	AdaBoostClassifier	0.806933	0.810945	0.725990	0.906280	0.545699
5	GradientBoostingClassifier	0.821689	0.800284	0.713577	0.897585	0.529570
6	XGBClassifier	0.925511	0.792466	0.702236	0.893720	0.510753

Fig: Accuracies of various machine learning algorithms applied

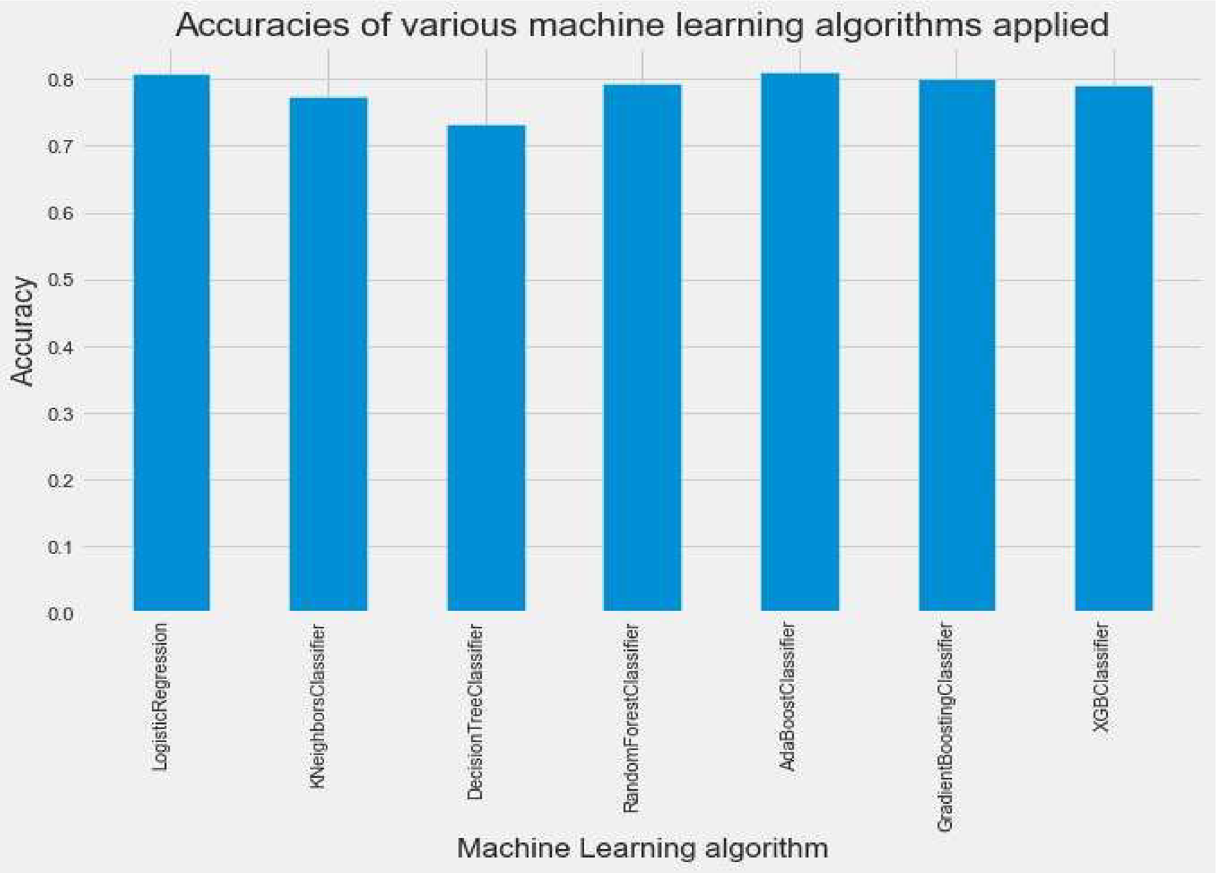


Fig: Accuracy of various machine learning algorithms applied as a bar chart

By above bar chart we got to know that Ada boost classifier was found to be most accurate algorithm and ROC curve and confusion matrix is displayed below.

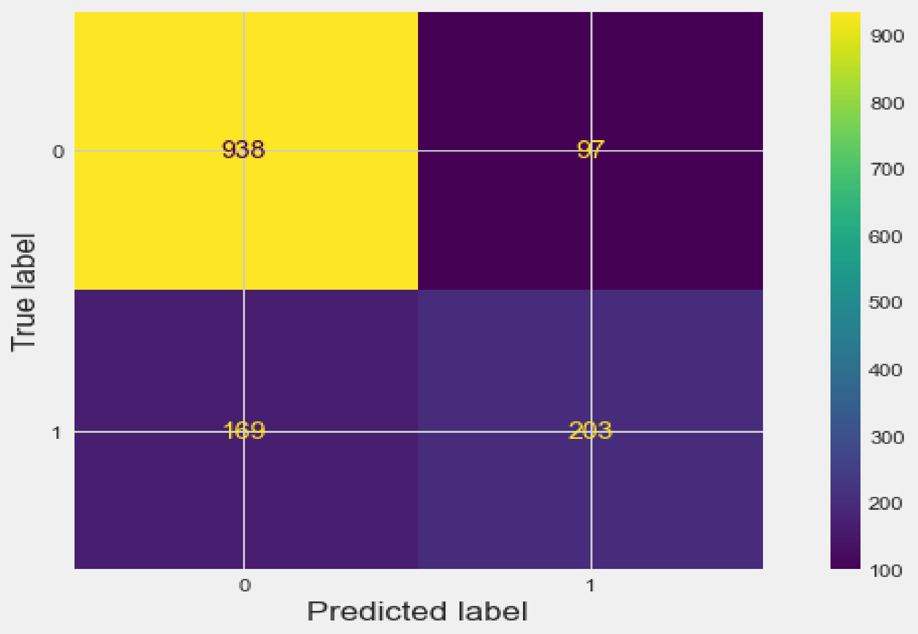


Fig : Confusion matrix for Ada Boost Classifier.

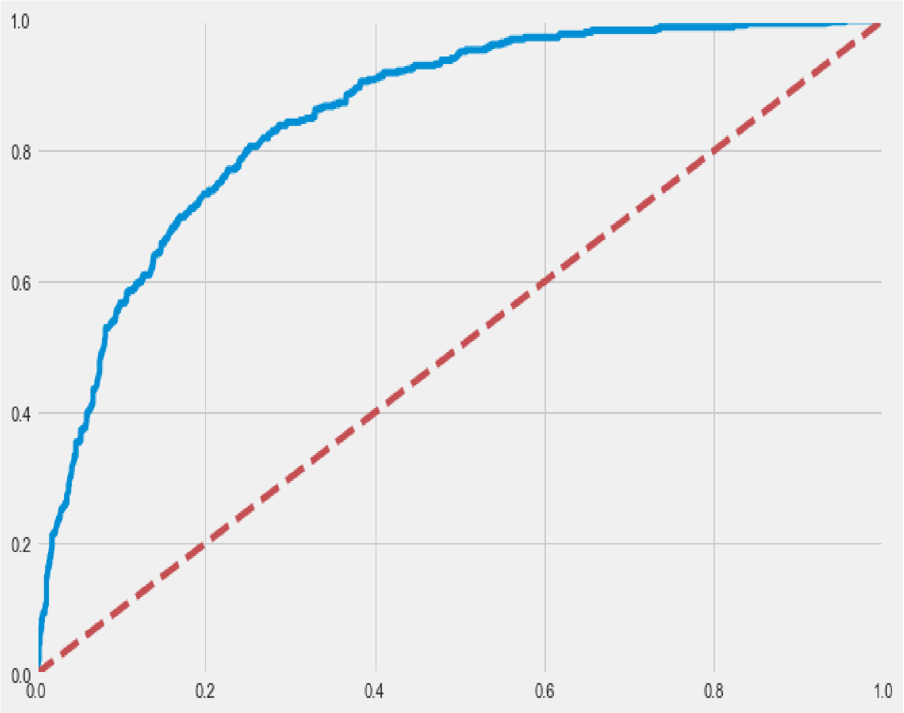


Fig : ROC curve for Ada Boost Classifier.

VI. CONCLUSION

The work that was proposed was mostly about using data mining techniques to do a comparative study on the extraction of features from a telecom dataset. With domain intelligence, the results from the attribute selection approaches can be used to find more specific attributes that help make good predictions. The data set is split into training data and testing data. Different machine learning algorithms are then used on the training data and their accuracy is compared. After comparing different algorithms, it was found that the Adaboost classifier algorithm was the most accurate, with a training accuracy of 80.69% and a testing accuracy of 81.9%.

REFERENCES

1] Rehman AS, Qamar AM, Kamal A, Rehman A, Qureshii SA, " Telecommunication subscribers' churn prediction model using machine learning" Eighth international conference on digital information management, 2013.
[2] Iyengar R, Schleicher M, Ascarza E, " The perils of proactive churn prevention using plan recommendations: evidence from a field experiment," J Market Res, 2016.
[3] Bott," Predicting customer churn in telecom industry using multilayer preceptron neural networks: modeling and analysis," Igarss, 2014.
[4] Iyakutti K, Umayaparvathi V, "A survey on customer churn prediction in telecom industry: datasets, methods, and metric," Int Res J Eng Technol, 2016.
[5] Jutla DN, Sivakumar SC, Yu W, "A churn-strategy alignment model for managers in mobile telecom," Communication networks and services research conference, 2005.
[6] Den Poel V, Burez D, "Handling class imbalance in customer churn prediction," Expert Syst Appl, 2009.
[7] Brandusoiu I, Toderean G, Ha B, "Methods for churn prediction in the prepaid mobile telecommunications industry," International conference on communications, 2016.
[8] He Z, Zhang D, He Y, "A study on prediction of customer churn in fixed communication network based on data mining," Sixth international conference on fuzzy systems and knowledge discovery, 2009.
[9] Huang F, Zhu M, Yuan K, Deng EO, "Telco churn prediction with big data," ACM SIGMOD international conference on management of data, 2015.
[10] Makhtar M, Zhu M, Yuan K, Deng EO, "Churn classification model for local telecommunication company based on rough set theory," J Fundam Appl Sci, 2017. [11] Gerpott TJ, Rams W, Schindler A "Customer retention, loyalty, and satisfaction in the German mobile cellular telecommunications market," Telecommun Policy, 2001.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

☎ 9940 572 462 6381 907 438 ✉ ijircce@gmail.com

Scan to save the contact details

www.ijircce.com