International Journal of Engineering & Technology, 7 (3.34) (2018) 291-295

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Comparative Analysis of Machine Learning Techniques to Identify Churn for Telecom Data

M.Malleswari1, Maniraj.R2, Praveen Kumar3, Murugan3

Department Of Information Technology,
Vel Tech High Tech Dr. Rangarajan, Dr. Sakunthala Engineering College.
**Corresponding author E-mail: praveen13996@gmail.com, muruganb3295@gmail.com* [2]

Abstract

Big data analytics has been the focus for large scale data processing. Machine learning and Big data has great future in prediction. Churn prediction is one of the sub domain of big data. Preventing customer attrition especially in telecom is the advantage of churn prediction. Churn prediction is a day-to-day affair involving millions. So a solution to prevent customer attrition can save a lot. This paper propose to do comparison of three machine learning techniques Decision tree algorithm, Random Forest algorithm and Gradient Boosted tree algorithm using Apache Spark. Apache Spark is a data processing engine used in big data which provides in-memory processing so that the processing speed is higher. The analysis is made by extracting the features of the data set and training the model. Scala is a programming language that combines both object oriented and functional programming and so a powerful programming language. The analysis is implemented using Apache Spark and modelling is done using scala ML. The accuracy of Decision tree model came out as 86%, Random Forest model is 87% and Gradient Boosted tree is 85%.

Keywords*: Churn prediction, Machine learning, Scala, Apache Spark, Big Data.*

1. Introduction

Big data is a about processing terabytes of data for business intelligence at real time. It is mainly used for performing analytics for prediction, recommendation, fraud detection and security monitoring and much beyond, hence data analytics is required at every stage, be it at run time or as well as at offline. It includes five types of analytics which are prescriptive analytics, descriptive analytics, diagnostic analytics, outcome analytics and predictive analytics. This project is based on predictive analytics— a technique which is used commonly which forecast what might happen under a specific situation using models. In predictive analytics the historical data, machine learning, and artificial intelligence are used to predict what happens in the future. The large set of historical data is fed into the mathematical model which considers the patterns in the data. One of the best example for predictive analytics is churn risk analytics— an analytics made to predict which customer is likely to leave or abandon the service. Churn prediction is especially used in telecom industry.
To satisfy the complex needs of telecommunication organizations a type of business intelligence is applied specifically and packaged which is known as telecom analytics. The focus of telecom analytics is at decreasing operational cost, maximizing profits, improving risk management and reducing fraud. Predicitive and descriptive modelling are used as well as forecast, optimization and multidimensional analyses are involved in telecom analytics. Fig 1 shows the analytics made in the telecom analytics.

2. Related Work

ammar a.q ahmed and maheswari d [1] proposed churn prediction on huge telecom data using hybrid firefly and swarm optimization, it predicts roc, pr, accuracy, true positive rate, false positive rate, f-measure and execution time. yiqing huang et al [3] predicts telco churn prediction using big data, it functions variety, volume and velocity,

Fig 1: Telecom Analytics

Prediction of customer involve of support system includes business support system and operating support system. scott a.neslin et al [9] proposed defection and detection techniques to measure and analysis churn prediction. some of the techniques to predict churn are of logistic regression, eda, and stepwise procedures for predictive analysis. theresa morelli et al [6] executes customer churn by applying ibm predictive customer intelligence. it solves customer queries to increase market share for the company to develop their average revenue per user and revenue generating unit. anuj sharma et al [7] framework a neural

network approach for predicting customer churn in telecommunication field. Artificial neural network technique will store data to avoid risk of churn; wireless network services will form an customer relationship management system for mass marketing strategies. it process fuzzy and non-linear to form pattern recognition to preform churn prediction in cellular network services. ajay chandramouly et al [5] proposed reducing client incidents through big data predictive analytics, it visualize extract, transform and to develop customer churning, comprises of event, sqoop, predict and import. junxiang lu [10] predict customer churn by an application of survival analysing modeling using sas, conventional statistical methods consists of logistics regression and decision tree, gains chart with sas code will implement the churn analysis in graphical manner. amoo a.o et al [2] perform modeling and simulation of a predictive customer churn model, it represents adaptive neuro fuzzy inference system to predict churning, exhaustive search algorithm with fuzzy rules will stimulate, validate and verified churn using matlab programming tools.

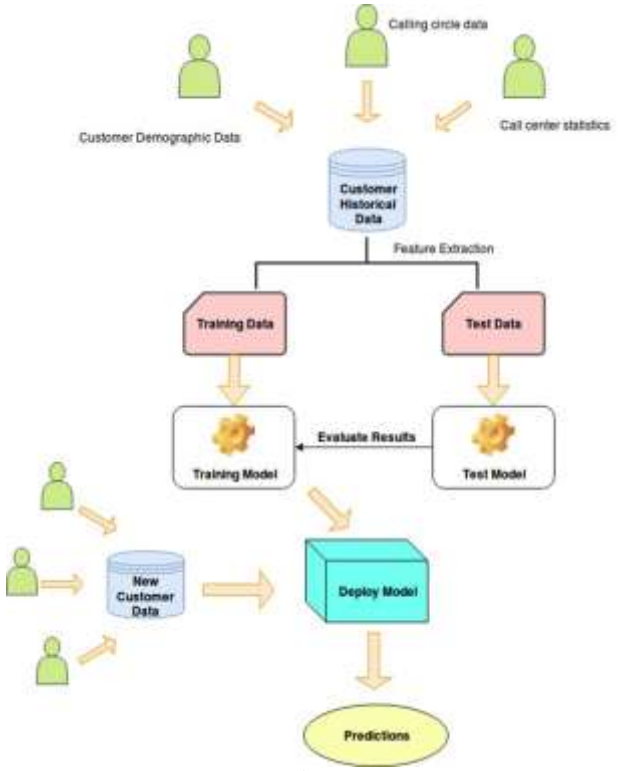

Fig 2: Architecture of churn prediction model

3. Comparative Analysis

The feature extraction is made from the data set and the prediction model is created using bagging and boosting classification algorithms. Bagging and Boosting algorithm— They are two ensemble techniques used to improve the accuracy of the classification tree algorithms. In ensemble technique n number of decision trees are generated based on the size of the training data set and the trees are combined to produce better predictive performance.

Bagging— is used to reduce the variance of the decision tree. Several subsets of data are created from the training data and each subset is used to train their decision trees. Each tree produces predictions and average of all the predictions is used to make final prediction. It is more robust when average of a number of trees used than a single decision tree. Random Forest algorithm is a great example for bagging algorithms.

Boosting— is also an ensemble technique where n number of predictors are created consecutively. The tree produced after the early tree learns sequentially by analyzing the data for errors. At each step, the aim is solving the net error from the previous tree. Gradient Boosted Tree algorithm is a very good example for boosting algorithms.

In this paper three classification algorithms are used to create predicition model and they are Decision Tree algorithm, Random Forest algorithm which comes under bagging and Gradient Boosted Tree algorithm which comes under boosting.

3.1. Architecture of Churn Prediction Model

Fig 2 shows the architecture of churn prediction model. In this paper the Orange Telecom data set is used for training the model.

Customer historical data is gathered which consist of the customer's state, account length, area code, international plan, voice mail plan, number of voice mail messages, total minutes per day, total calls per day, total charge per day, total evening minutes, total evening calls, total evening charge, total night minutes, total night calls, total night charge, total international minutes, total international calls, total international charge, customer service calls.

Feature extraction is the process of collecting features that are necessary and removing the unwanted features from the dataset. The cost of a call can be derived from the minutes of the call so such correlated features (total day charge, total day minutes) are removed. The features state, area code, voice mail plan are not used as well so these features are also removed.

The whole data set is split into testing and training data set where the testing data set contains a 20% and training data set contains 80% of the whole data set. The ratio of training data should be higher in order to produce a model with better accuracy.

Training data consists of features with their values respectively and label. In this paper, label is the title that contains the information whether the customer is churned or not churned. Testing data also consists of features and label which is used to evaluate the accuracy of the model after training. After evaluation the new customer data is used in the model and the customer churn is predicted.

3.2. Decision Tree Algorithm based Prediction

Decision trees create a prediction model that predicts the label (churn), based on several input features. Decision trees evaluates an expression containing a feature at every node and based on the result it selects the branch to the next node.

3.3. Trained Decision Tree Model

Here the feature 1 to feature 11 are international plan, number of voice mail, total day minutes, total day calls, total evening minutes, total evening calls, total night minutes, total night calls, total international minutes, total international calls, and number of customer service calls.

DecisionTreeClassificationModel of depth 5 with 53 nodes
If (feature 11 <= 3.5)
If (feature 3 <= 222.7)
If (feature 1 in {1.0})
If (feature 9 <= 13.149999999999999)
If (feature 10 <= 2.5)
Predict: 0.0
Else (feature 10 > 2.5)
Predict: 1.0
Else (feature 9 > 13.149999999999999)
Predict: 0.0
Else (feature 1 not in {1.0})
If (feature 4 <= 125.5)
If (feature 3 <= 209.55)
Predict: 1.0
Else (feature 3 > 209.55)

Predict: 1.0
Else (feature 4 > 125.5)
If (feature 3 <= 161.25)
Predict: 1.0
Else (feature 3 > 161.25)
Predict: 0.0
Else (feature 3 > 222.7)
If (feature 2 <= 6.0)
If (feature 5 <= 183.95)
If (feature 3 <= 273.35)
Predict: 1.0
Else (feature 3 > 273.35)
Predict: 0.0
Else (feature 5 > 183.95)
If (feature 3 <= 242.25)
Predict: 0.0
Else (feature 3 > 242.25)
Predict: 0.0
Else (feature 2 > 6.0)
If (feature 1 in {1.0})
If (feature 0 <= 57.5)
Predict: 1.0
Else (feature 0 > 57.5)
Predict: 0.0
Else (feature 1 not in {1.0})
If (feature 3 <= 299.9)
Predict: 1.0
Else (feature 3 > 299.9)
Predict: 0.0
Else (feature 11 > 3.5)
If (feature 3 <= 181.0)
If (feature 10 <= 0.5)
Predict: 1.0
Else (feature 10 > 0.5)
If (feature 5 <= 273.1)
If (feature 5 <= 226.14999999999998)
Predict: 0.0
Else (feature 5 > 226.14999999999998)
Predict: 0.0
Else (feature 5 > 273.1)
If (feature 0 <= 99.5)
Predict: 0.0
Else (feature 0 > 99.5)
Predict: 1.0
Else (feature 3 > 181.0)
If (feature 8 <= 104.5)
If (feature 8 <= 81.5)
If (feature 4 <= 94.5)
Predict: 1.0
Else (feature 4 > 94.5)
Predict: 0.0
Else (feature 8 > 81.5)
If (feature 4 <= 137.5)
Predict: 1.0
Else (feature 4 > 137.5)
Predict: 0.0
Else (feature 8 > 104.5)
If (feature 0 <= 135.5)
Predict: 0.0
Else (feature 0 > 135.5)
If (feature 0 <= 160.5)
Predict: 1.0
Else (feature 0 > 160.5)
Predict: 0.0

The decision tree makes a decision based on the above trained model.

3.4. Accuracy

True positives are the positives that the model predicted correctly. False positives are the positives that the model predicted wrongly. True negatives are the negatives that the model predicted correctly. False negatives are the negatives that the model predicted wrongly. Fig 3 represents the accuracy of decision tree model on predicting positive churn of the customer.

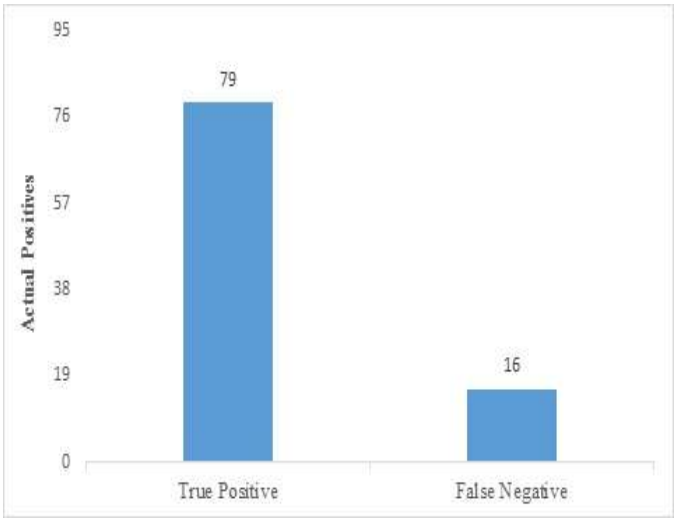

Fig 3: Positive Churn Prediction Accuracy

Fig 3 represents the accuracy of decision tree model on predicting positive churn of the customer.
The decision tree model predicted 79 positives correctly and 16 positives wrongly out of 95 positives.

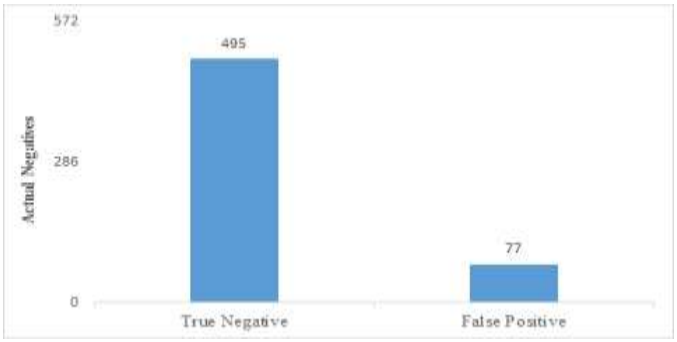

Fig 4: Negative Churn Prediction Accuracy

Fig 4 represents the accuracy of decision tree model on predicting negative churn. The decision tree model predicted 495 negatives correctly and 77 negatives wrongly out of 572 negatives.

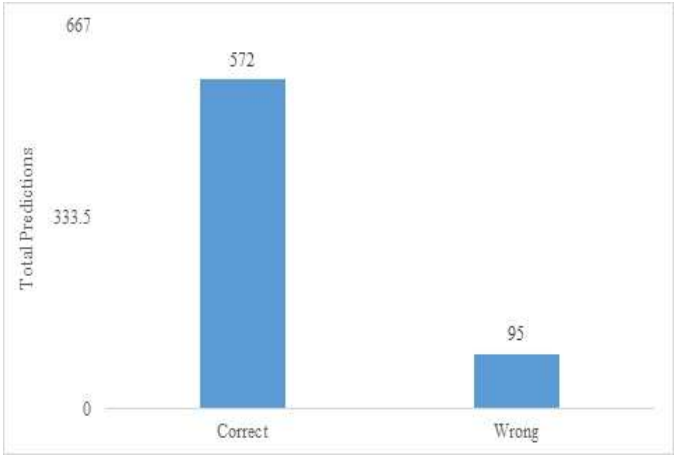

Fig 5: Total Accuracy

Fig 5 represents the accuracy of decision tree model. The decision tree model predicted 574 customer churns correctly and 73 customer churns wrongly out of 667 customers. Thus the accuracy of decision tree model is 86%.

3.5. Random Forest Algorithm based Prediction

Random forest algorithm is a classification algorithm that generates n number of decision trees parallely. The number of trees generated is based on the size of the training data set. The decision trees generated are trained with different parts of the training data set. Each tree produces their prediction and the average of the predictions is used to make the final prediction.

3.6. Accuracy

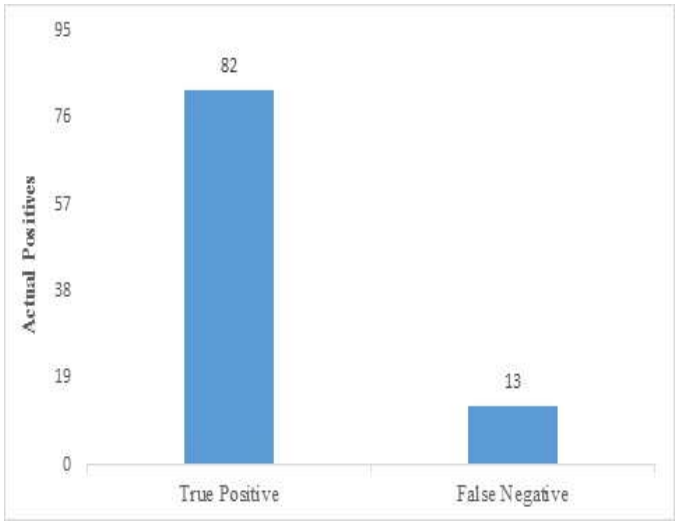

Fig 6: Positive Churn Prediction Accuracy

Fig 6 represents the accuracy of random forest model on predicting positive churn of the customer. The random forest model predicted 82 positives correctly and 13 positives wrongly out of 95 positives.

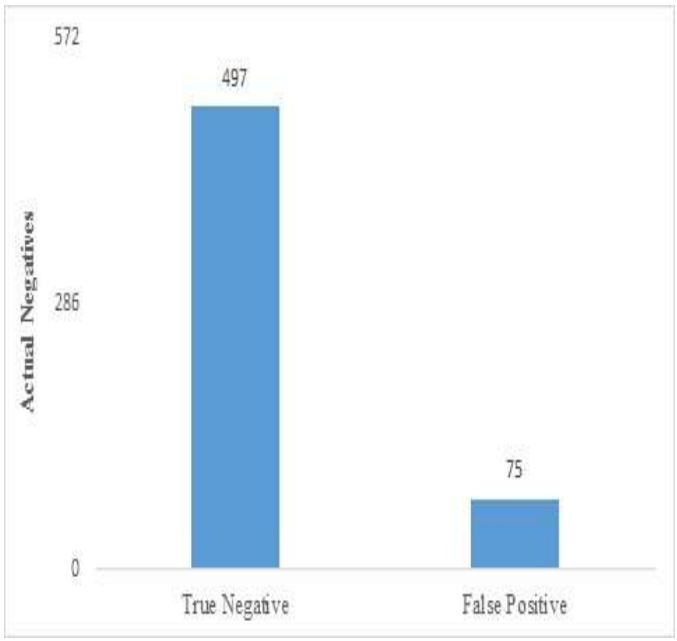

Fig 7: Negative Churn Prediction Accuracy

Fig 7 represents the accuracy of random forest model on predicting negative churn of customer. The random forest model predicted 497 negatives correctly and 75 negatives wrongly out of 572 negatives.

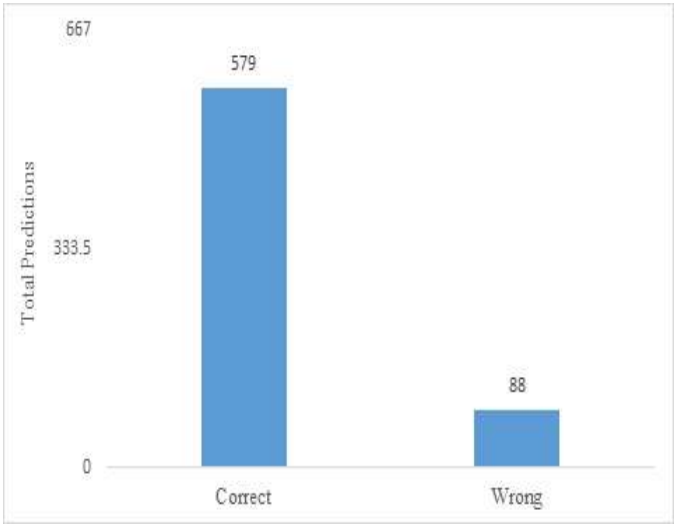

Fig 8: Total Accuracy

Fig 8 represents the total accuracy of random forest model**.** The random forest model predicted 578 customer churns correctly and 88 customer churns wrongly out of 667 customers. Thus the accuracy of random forest model is 87%.

3.7. Gradient Boosted Tree Algorithm based Prediction

Gradient Boosting algorithm is a boosting algorithm which generates n number of decision trees subsequently. Like random forest algorithm the number of decision trees generated is based on the size of the training data set. The decision trees generated depends upon the result of its previous tree. The generated tree learns the mistakes of the previous tree by analysing the data for error and solves the net error from the previous tree. The execution time taken by this algorithm is higher as the trees are generated sequentially.

3.8. Accuracy

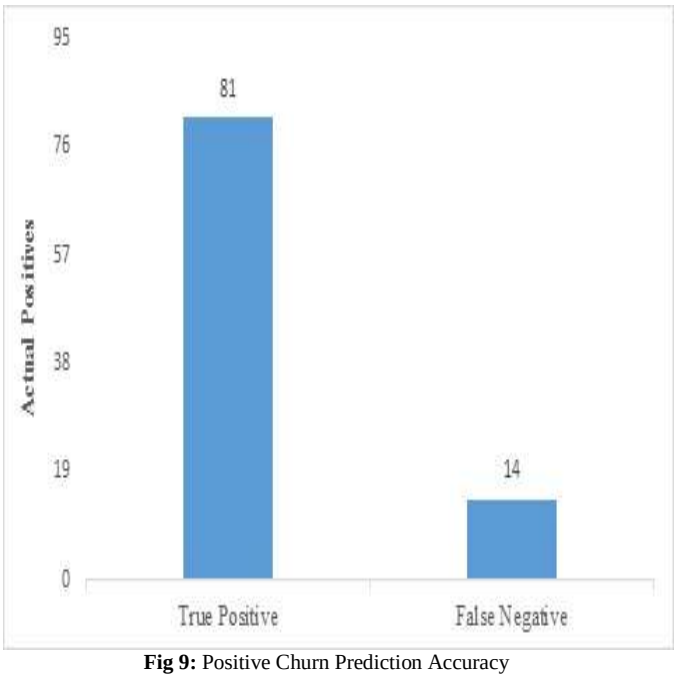

Fig 9: Positive Churn Prediction Accuracy

Fig 9 represents the accuracy of gradient boosted tree model on predicting positive churn of customer. The gradient boosted tree model predicted 81 positives correctly and 14 positives wrongly out of 95 positives.

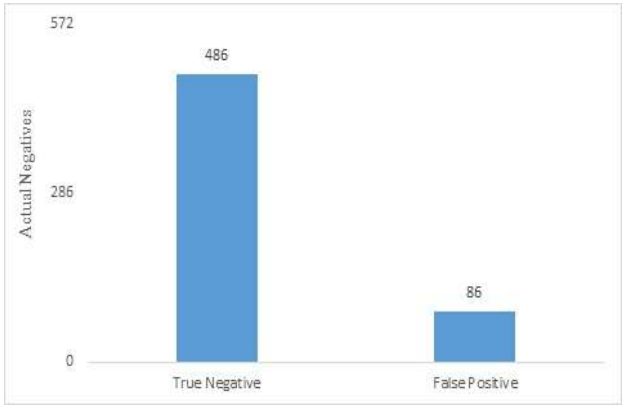

Fig 10: Negative Churn Prediction Accuracy

Fig 10 represents the accuracy of gradient boosted tree model on predicting negative churn of customer. The gradient boosted tree model predicted 486 negatives correctly and 86 negatives wrongly out of 572 negatives.

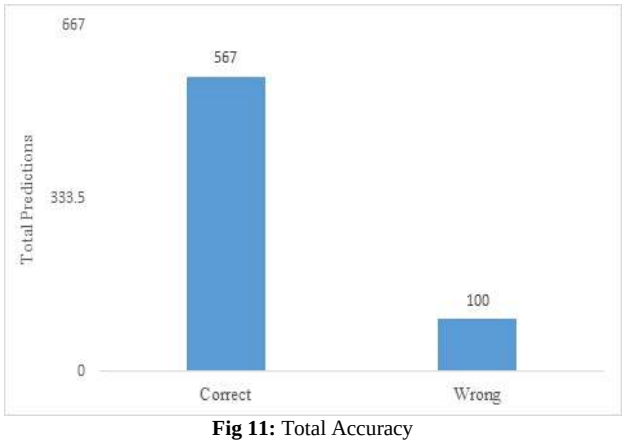

Fig 11: Total Accuracy

Fig 11 represents the total accuracy of gradient boosted tree model. The gradient boosted tree model predicted 567 customer churns correctly and 100 customer churns wrongly out of 667 customers. Thus the accuracy of gradient boosted tree model is 85%

3.9. Comparison of the accuracy of Decision Tree, Random Forest and Gradient Boosted Tree Models

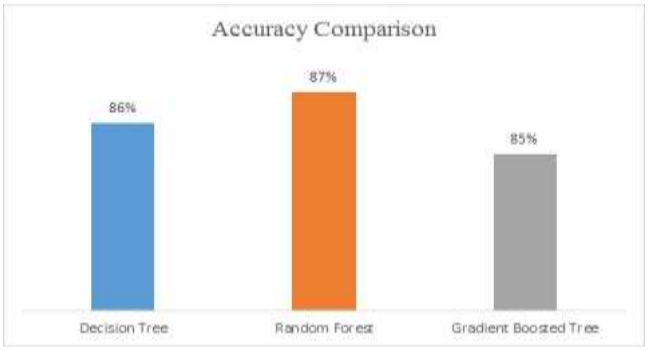

Fig 12: Comparison

The accuracy of random forest algorithm is higher than the other two algorithms.

4. Conclusion

Churn prediction has become very important in telecom industry due to the increase in number of customers day by day. Customer retention practices should be made by the industry to make the customer stay in the subscription who is likely to cancel the subscription.

This paper is involved around this issue. It looks to solve this problem with an efficient solution. We compared three machine learning algorithm's accuracy on churn prediction. It is concluded that the accuracy of the Gradient Boosted Tree algorithm is lower than the accuracy of other two algorithm. The execution time of the Gradient Boosted Algorithm is also higher than the execution time of other two algorithms. But the execution time of Decision Tree algorithm is lower than the execution time of other two algorithms. Even though the execution time of Random Forest algorithm is higher than Decision Tree algorithm, the accuracy of Random Forest algorithm is better than the other two algorithm.

References

[1] Ammar A.Q Ahmed, Maheswari D "Churn Prediction on Huge Telecom Data Using Hybrid Firefly- Particle Swarm Optimization Algorithm Based Classification" IOSR Journal of Computer Engineering (IOSR-JCE)e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 4, Ver. VII (Jul.-Aug. 2017), PP 30-39

[2] Amoo A. O, Akinyemi B. O, Awoyelu I. O, Adagunodo E. R, "Modeling & Simulation of a Predictive Customer Churn Model for Telecommunication Industry" Vol. 6, No. 11, November 2015 ISSN 2079-8407 Journal of Emerging Trends in Computing and Information Sciences.

[3] Yiqing Huang, Fangzhou Zhu, Mingxuan Yuan, Ke Deng, Yanhua Li, Bing Ni, Wenyuan Dai, Qiang Yang, Jia Zeng "Telco Churn Prediction with Big Data" SIGMOD'15,May31–June 4, 2015, Melbourne, Victoria, Australia.Copyright c 2015 ACM 978-1-4503-3469.

[4] Dr.M.Balasubramanian, Dr.M.Selvarani, "Churn Prediction In Telecom System Using Data Mining Techniques", International Journal of Scientific and Research Publications, Volume 4, Issue 4, April 2014, ISSN 2250-3153.

[5] Ajay Chandramouly, Ravindra Narkhede, Vijay Mungara, Guillermo Rueda, Asoka Diggs, "Reducing Client Incidents through Big Data Predictive Analytics" Intel IT IT Best Practices Big Data Predictive Analytics December 2013.

[6] Theresa Morelli, Vivian Braun, David Pugh, Venky Rao, "Retain and Delight Your Customers by Applying IBM Predictive Customer Intelligence" Empowered Customers Drive Collaborative Business Evolution®, Forrester Research, Inc, May 2012.

[7] Anuj Sharma, Dr. Prabin Kumar Panigrahi, "A Neural Network based Approach for PredictingvCustomer Churn in Cellular Network Services" International Journal of Computer Applications (0975 – 8887)Volume 27– No.11, August 2011.

[8] Rahul J.Jadav, Usharani T.Pawar, "Churn Prediction in Telecommunication Using Data Mining Technology", International Journal of Advanced Computer Science and Applications, Vol. 2, No.2, February 2011.

[9] Scott A. Neslin, Sunil Gupta, Wagner Kamakura, Junxiang Lu, and Charlotte H. Mason, "Defection Detection: Measuring and understanding the Predictive Accuracy of customer churn models" Journal of Marketing ResearchVol. XLIII (May 2006),204211204© 2006, American Marketing Association.

[10] Junxiang Lu, "Predicting Customer Churn in the Telecommunications Industry — An Application of Survival Analysis Modeling Using SAS" SAS Institute Inc., 2001.

[11] P. Datta, B. Masand, D. Mani, and B. Li. Automated cellular modeling and prediction on a large scale. Artificial Intelligence Review, 14(6):485–502, 2000.

[12] M. Wegmuller, J. P. von der Weid, P. Oberson, and N. Gisin, "High resolution fiber distributed measurements with coherent OFDR," in Proc. ECOC'00, 2000, paper 11.3.4, p. 109.

[13] S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," IEEE Electron Device Lett., vol. 20, pp. 569–571, Nov. 1999.

[14] S. M. Metev and V. P. Veiko, Laser Assisted Microtechnology, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.

[15] R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, "High-speed digital-to-RF converter," U.S. Patent 5 668 842, Sept. 16, 1997.