International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 04 | Apr 2021 www.irjet.net p-ISSN: 2395-0072

CHURN ANALYSIS FOR MAXIMIZING PROFIT USING STATISTICAL MODEL IN TELECOM INDUSTRY

G.Suguna[1]

PG Scholar, Department of Computer Science and Engineering, University College Of Engineering, Tamilnadu, India

---

***Abstract* —** Churn is a vital business metric for subscription-based services like telecommunication corporations. A telecommunication company needs a predictive model for finding the churning of customers in telecom industry. The predictive model should categorize the customers likely to churn and the customers who settle in the industry. The built model should avoid the waste of resources. The cost related with getting a new customer is 8 times greater than the existing customer. The target is to look for general patterns within the provided information that might facilitate to spot customers that square measure a lot of possible to churn, before doing that, the corporate may address them and stop their churn. This paper demonstrates a churn analysis tested LOGISTREE MODEL which combines Logistic regression and decision tree for predicting churn. It consists of two stages: a segmentation section and a prediction section. In the first stage client segments square measure known tested call rules and in the second stage a model is formed for each leaf of this tree. This proposed hybrid approach is benchmarked against logistic regression, decision tree with regards to the predictive performance and quality.

Keywords—Churn Prediction; Logistic Regression; Decision Trees; Client Retention

I. INTRODUCTION

The telecommunication sector has grown-up as one among the important businesses in urbanized countries. The technological advancements and the rising operators increased the degrees of competition. Organizations are working very hard to survive this competition.

Three strategies have been planned to make a lot of revenues like obtaining new customers, up sell the on hand customers, to increase the period of retention time of customers. The third procedure is the most profitable technique, it demonstrates that holding a current consumer needs less cost than acquiring a new customer. To employ the third strategy, organizations must reduce the potential of client's churn.

Customer's churn is the main concern in administration sectors with high competitions. Predicting the consumer who are expected to leave the organization will correspond to actually large additional revenue source if it is done in the in the initial hours of the phase. Several analysis suggests that machine learning is extremely fruitful to anticipate this circumstance.

Data mining techniques helps the telecom industry to use effective methods for minimizing the churn rate of customers. Early identification of customers who are susceptible to churn might help the telecom sector to increase their productivity. Inaccurate predictions may cause a company to lose profit.

This lead the companies to put a more effort in analyzing and understanding the customers' behavior, in order to identify with adequate advance which customer will leave. In particular, these customers are subjected to several rectifying actions like gifts or promotions for their retention.

II. RELATEDWORK

Most of the analysis in telecommunication industry for prediction of churned or non-churned for retaining the customers. To create a model based on the prediction with the help of particular classification of customers with the help of historical data for retaining the customer from churning.

Hemlata Jain et al.,[1] presented **"Churn Prediction in Telecommunication using Logistic Regression and Logit Boost"** mentioned regarding corporations proactively ought to verify the purchasers churn by analyzing their actions and take a look at to place effort and cash in holding the purchasers. During this planned model, two machine-learning techniques were used for forecasting cash in holding the purchasers. Logistic regression and Logit Boost were used for forecasting client churn.

Arifin, S et al., [2] presented **"Analysis of churn rate significantly factors in telecommunication industry using support vector machines method"** This analysis was supposed to understand features that direct

the churn rate considerably in telecom industry through analysis of historical request and identification information of the customers. This work has totally seven variables, four of them were historic variables and three of them was identification variables. This information was taken from the telecom industry in Dutch East Indies by taking active user with minimum six months and historical data from Gregorian calendar month till March 2018. The data was tested by using Support Vector Machine by taking the results of classification performance either on the performance of all the variables or performance of individual variable.

Choudhari, A. S et al., [3] presented **"Predictive to prescriptive analysis for customer churn in telecom industry using hybrid data mining techniques".** This article focusses on knowledge processing approach for predicting the churn in telecommunication business, which face significant loss of financial gain, due to shopper in peril of leaving from an organization. The proposed model will be able to predict churn pattern of the subscriber well in earlier. They have projected hybrid Classification techniques that have revealed their dominances over single technique. The approach is to match two styles of hybrid strategies, with classification and cluster along with classification hybrid strategies like Fuzzy unordered induction rule with Fuzzy C-means cluster for forecasting the churn of the client.

Rajamohamed, R et al., [4] presented **"Improved credit card churn prediction based on rough clustering and supervised learning techniques"** ideas concerning Customers retaining technique in credit card churn prediction was performed by supervised classification techniques. However it couldn't yield higher results. They used several verified hybrid classification techniques which gives higher accuracy in C3P. Also C3P lags in extremely economical techniques like rough pure mathematics. During this work at first, it has a tendency to perform processing techniques and in the next stage, a tendency to propose changed rough K-means rule used for clustering the credit card holders and in next stage hold-out methodology splits the cluster information into testing and training sets. Finally grouping is performed and tested for several algorithms like random forest, support vector machine, call tree, K-nearest neighbor, and Naive Bayes. It has a tendency to define the work using measures like accuracy, recall, specification and misclassification error.

Saran Kumar, A et al., [5] presented **"A Survey on Customer Churn Prediction using Machine Learning Techniques"** during this study mentioned concerning the quick growth of the market in each sector is resulting in superior subscriber base for service suppliers, more competitors, novel and innovative business models and increased services ar increasing the value of client acquisition. The quick growth of the market in each sector is resulting in superior subscriber base for service suppliers.

Saghir. M et al., [6] presented **"Churn prediction using neural network based individual and ensemble models"** this paper targets to develop an understandable approach using data mining, so as to attain 2 goals: first is to produce an extremely correct and sturdy demand prediction model of re-manufactured products and second is to shed light-weight on the nonlinear result of on-line market factors as prognosticators of client demand that supports the real world Amazon dataset. The planned method will predict the merchandise demand with high accuracy.

Bi, W., Cai, M et al., [7] presented **"A big data clustering algorithm for mitigating the risk of customer churn"** discussed concerning several data processing techniques are utilized to forecast the churn of the customer and therefore scale back churn rate. Through various algorithms there is scope for improving the performance. This work assesses prevailing individual associated Neural Network ensemble using classifiers and recommends an ensemble classifier that employs stacking with Neural Network so as to enhance performance measures leading to higher precision for predicting the churn of customers.

Brandusoiu I et al., [8] presented **"Methods for churn prediction in the prepaid mobile telecommunications industry"** In this paper, there is tendency to build a complex data processing methodology that predicts client churn within the prepaid mobile communications business employing a call log dataset of 3333 customers with twenty one attributes. In this work the principal part analysis is applied first to reduce the high dimensionality. A technique using machine learning algorithms like support vector machines, neural networks, and Bayes network to analyze the models.

Jayaswal, P et al., [9] presented **"An Ensemble Approach for Efficient Churn Prediction in Telecom Industry"** This work focuses the telecommunication Service suppliers for analysis of Churned and Non-churned customers. In this analysis they have used Gradient Boost, RFand Ensemble classifiers for classification of churned and non-churned customers. The evaluation has been done with

the apache spark and it is based on unified knowledge analysis for pre- processing the data. To obtain the result they have used optimization with hyper parameter.

Idris A et al., [10] presented **"Churn prediction system for telecom using filter–wrapper and ensemble classification"** presented a new improvised churn prediction algorithm for telecommunication services such as FW ECP. FW ECP has the ability to combine each filter associated with the wrapper based on features selection and enhance the training ability of the classifier and was tested with various base classifiers. Particle Swarm Optimization is employed for under sampling in the filter phase, for feature selection they have used mRmR technique. The Wrapper phase uses Genetic algorithm for removing the identical and irrelevant features. RF2, RBoost, SVM are utilized for taking advantage of the new selection of features.

Anjum, A et al., [11] presented **"Optimizing Coverage of Churn Prediction in Telecommunication Industry"** this work proposes a decision support system for predicting the churning behavior of a client. Analytical system was developed by using machine learning techniques like CHAID, C5 and ANN and QUEST for the churn analysis and prediction for the telecommunication business. Prediction performance is considerably improved by employing a giant volume and several other options from each Business Support Systems and Operations Support Systems Detailed experiments were done to improve the performance. Significant increase in predictive performance was obtained.

Azeem, M et al., [12] presented **"A fuzzy based churn prediction and retention model for prepaid customers in telecom industry"** a distinct churn prediction and retention model for attaining the exact identification and expected retention of churners. They have done correct churn identification at different levels of severity by employing fuzzy classifiers. The model mechanically produces smart retention promotions by mining client usage and complaints patterns. Moreover the built model produces smart retention operations automatically. They have produced 98% accuracy of churned class using real telecom data set. Moreover, the planned strategy and categorization retained 87% of the potential churned customers.

Ahmad A.K et al., [13] presented **"Customer churn prediction in telecom using machine learning in big data platform"** The main intention of the analysis is to check concerning churn prediction of the client in telecommunication tested machine learning in massive knowledge platform. Machine learning approaches play a significant role in predicting the buyer churn. This analysis makes use of KNN with massive knowledge for predicting the buyer churn within the telecommunication business. From the findings of the result, it had been found that accuracy rate of prediction in client churn is found to be 0.80 % and space below curve is found to be 0.71 %.

III. PROPOSEDSYSTEM

In proposed system the analysis of performance measures for the created model is based on Decision tree(call tree) and Logistic regression, illustrated with confusion matrix analysis to retain the customers. The user should be aware of the dataset and the features which are selected based on the results obtained whether the customers are churned or non-churned. The user should desire to use the particular features for his or her own knowledge of various formats and the testing options for the particular classification method. The created model is focused on the deeper classification about the customers based on the services which are used by them. From the results of the model the telecommunication industry can enlarge the services and to retain the old customers.

This analysis is build with the help of data analytic tool i.e., R Programming. This tool is available as open source and mainly for mining knowledge about the data, applied for math computer code analysis, used among many statisticians for knowledge analysis. In this research, R tool is used to predict the churned and non-churned customers. The features are given as input and explored the analysis with the specified attributes to create a model with decision tree and logistic regression to classify the customers. Dataset is split into testing and training data to generate a churned model based on the selected features from the dataset. The training data is used to create a model and testing set is mainly focused on evaluation of the model's performance.

This analysis about churn prediction in telecommunication industry is diagrammatically depicted with Decision trees are called classification trees or regression trees which supports the target variable. The classification with logistic regression is based on For regression trees, the values of the target variable fit into a contiguous domain. The decision trees are used with structure that is composed of consequences and nodes.

This method is simple to infer due to the way of representation of the output in the tree. Moreover, decision trees will handle different categories of input like numerical, categorical, or both. The performance of call trees is less

once it encompasses capturing advanced random relationship between the inputs and outputs.

3.1 Exploring the patterns in dataset tested call Tree

The most advanced part of a call tree is crucial, however the tree has to be designed. The primary objective is to separate the sample dataset into 2 subsets: a training set and testing set. The training dataset is important to create the call tree. The testing data set is tested against the model and the accuracy is recorded. Advantage of call trees is that they can be applied to all or any styles of knowledge and datasets which need very little cleansing and groundwork before it is accustomed to construct the model. Once the data is complete, the model is built. The process starts with the complete training dataset. The dataset must be divided into 2 smaller subsets to form branches of the tree. The most important goal is to form buckets that is pure as possible with respect to the target variable 'Churn'. The pure bucket, all customers would churn or all customers wouldn't. varied measures for purity measure accessible and will be used for call trees. Entropy is the measure used for telecom data.

The formulation of *entropy* is as follows:

$$Entropy\ (p) = -\sum_{i=1}^{N} p_i \log_2 p_i \text{ ---------- (1)}$$

P_i - proportion of property i within the set. This equation 1 is applied to the estimated variable -Churn. This has 2 values, zero and one. No churn is represented by and churn is represented by one. The purity of the bucket is measured by entropy zero and a bucket which has specifically half of the consumers would churn and half would not churn.

The goal of the call tree is to get buckets that leads the entropy closer to zero. Once the entropy is nearer to zero, it is easy to decide the client would churn or not and the client is located in the particular bucket. we calculate the purity of a bucket. The method to find out the best way is to divide the data into two. The Information gain of every split is calculated wherever it is possible.
The mathematical formula is:

$$IG(Y,X) = E(Y) - E(Y|X) \text{ -------- (2)}$$

Subtract the entropy of Y given X from the entropy of simply Y. For calculating the minimization of uncertainty about Y given an extra piece of data X regarding Y. It is called as Information Gain. If the Information Gain is more than the uncertainty is reduced. The split which has largest Information Gain is selected.

3.2 Fitting Logistic regression to the training set

Logistic Regression is used for binary classification problem. It predicts the chance of an event by finding the connection between a variable and one or more predictor variables (features). Logistic regression can also predict the probability of dependent variable. The condition for logistic regression is that the dependent variable should be binary.

Regarding the client churn, the chance of client churn occurrence. i.e., If y = 1, if churn occurs; else y = 0. The 2 conditional chances, p(y = 1|X) and p(y = 0|X) wherever X = (X_1, X_2...X_n) represents n factors(distinct variables) that are related to the client churn, are the possibility that a client goes to not churn or churn. This can be the result of the logistic regression model. the target is reducing the squared error of the classifier, that can be defined as follows:

$$\log\frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \text{ -------- (3)}$$

Where,

β_i = 1, ...n represents the coefficients of logistic regression.

β_0 is the intercept.

Xi, i = 0, 1, ...,n. are the predictor elements. These coefficients square measure determined as a part of the supply multivariate analysis. The above equation is often written as:

A proper threshold value δ has to be set. Once the threshold is decided, it may be inferred that a client is probably going to churn if p(y = 1|X) > δ . Applying regression has the ability of differentiating independent variables that square measure statistically important to have an effect on the output (customer churn).

The Sigmoid function is used for mapping the predictions to the probabilities. The function of sigmoid function is to compress the value of output from zero to one.

$$P(Y|X) = \frac{1}{1+\exp(-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n))} \text{ -------- (4)}$$

Applying regression is the best option amongst other preferred algorithms compared to classifier.

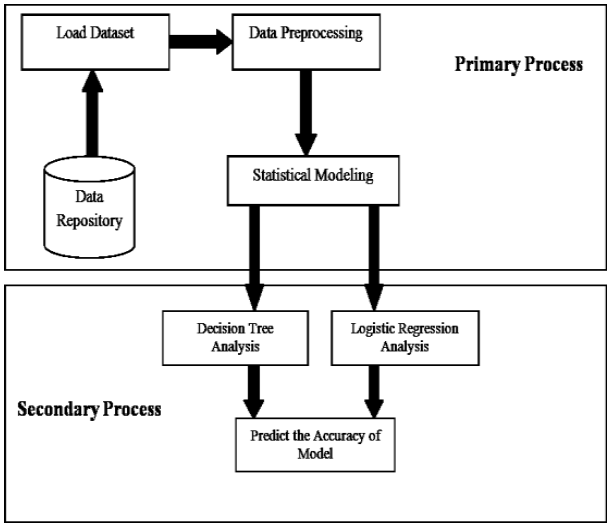


Figure 1: System Architecture

Many real world data is dirty and needs to be cleaned before employed in code. The method of cleanup a dataset is termed as knowledge pre-processing. Pre-processing of information includes below steps:

1. Taking care of missing knowledge
2. Categorical knowledge
3. Splitting dataset into test set and training set
4. Feature scaling

3.3 Fitting Random Forest to the training set

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

Working of Random Forest Algorithm

We can understand the working of Random Forest algorithm with the help of following steps –

- **Step 1** – First, start with the selection of random samples from a given dataset.
- **Step 2** – Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
- **Step 3** – In this step, voting will be performed for every predicted result.
- **Step 4** – At last, select the most voted prediction result as the final prediction result.

3.4 Exploratory Data Analysis

Exploratory knowledge Analysis (EDA) is that the method of analyzing and visualizing the information to urge a higher understanding of the information and collect insight from it.

There are numerous steps concerned once doing EDA however the subsequent are the common steps that a knowledge analyst will take once play acting EDA:

1. Import the data
2. Clean the data
3. Process the data
4. Visualize the data

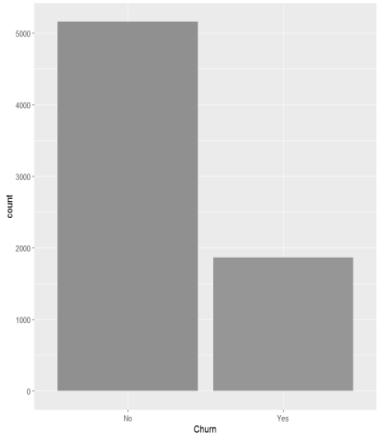


Figure 2: Looking at the proportion of customers that churned

Box plot for Continuous variable:

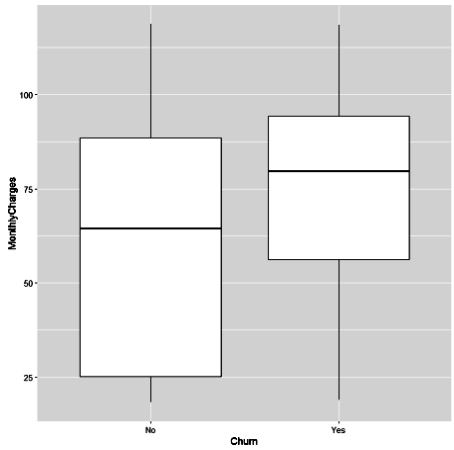


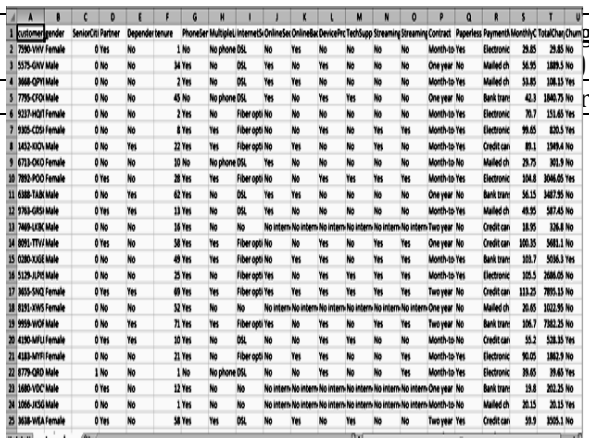
Figure 3: Pattern Exploration for the feature of monthly charges

3.5 Data Assortment

Data assortment is that the systematic approach to gathering and measurement data from a range of sources to induce a whole and correct image of a locality of interest. knowledge assortment permits someone or organization to answer relevant queries, valuate outcomes and create predictions concerning future chances and trends. correct knowledge assortment is crucial to maintaining the integrity of analysis, creating informed business choices and making certain quality assurance.

Table 1:Dataset Description

Sl. No	get s
1.	r 1



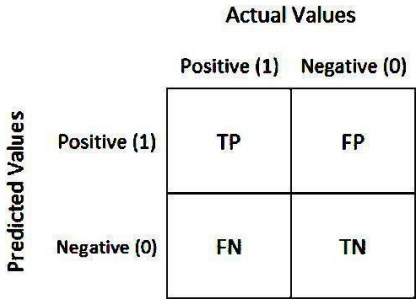
The dataset was downloaded from IBM Sample Knowledge Sets for client retention programs. The goal of this paper is to predict the behaviors of customers, as churn or not-churn. Every row denotes a customer; every column denotes a customer's attribute.

3.6 Evaluation

3.6.1 Confusion matrix:

A Confusion matrix is used to assess performance of the classification model. Confusion matrix relates the desired values with the actual values predicted by the machine learning model. It gives us an knowledge regarding the performance of the model. For a binary classification problem, 2 x 2 matrix is given with 4 quadrants.

Table 2: General formation of Confusion Matrix



Where,

True positive - predicted and actual values considered are true

False Positive - The actual value is false , but it is predicted as true

False Negative - The actual value is true, but it is predicted as false .

True Negative - predicted and actual values are considered as false

Most of the attributes with measures are used to assess the model created during the analysis. Confusion matrix was generated for every model and the matrix contains the desired values and actual values. Confusion matrix consists of True positive, False positive, False negative and True negative.

The following confusion matrix is calculated using LR:

```
cm_LR <- table(test_set[,18], y_pred_LR)
cm_LR
    y_pred_LR
        0    1
No   1157  134
Yes   237  230
```

```
> example <- test_set[2,]
> example
  gender SeniorCitizen Partner Dependents PhoneService MultipleLines
7   Male            No      No         Yes          Yes           Yes
  InternetService OnlineSecurity OnlineBackup DeviceProtection TechSupport
7     Fiber optic             No           Yes               No          No
  StreamingTV StreamingMovies       Contract PaperlessBilling
7         Yes              No Month-to-month              Yes
            PaymentMethod MonthlyCharges Churn tenure_group
7 Credit card (automatic)           89.1    No  12-24 Months
>
> example_prob <- prob_pred_LR[2]
> example_prob
        7
0.4494433
>
>
> example <- test_set[8,]
> example
   gender SeniorCitizen Partner Dependents PhoneService MultipleLines
32   Male            Yes     Yes         No          Yes            No
   InternetService OnlineSecurity OnlineBackup DeviceProtection TechSupport
32     Fiber optic             No           No               Yes          No
   StreamingTV StreamingMovies       Contract PaperlessBilling
32         Yes              Yes Month-to-month              Yes
             PaymentMethod MonthlyCharges Churn tenure_group
32 Credit card (automatic)           95.5    No  0-12 Months
>
> example_prob <- prob_pred_LR[8]
> example_prob
       32
0.7091426
```

Accuracy is calculated using the formula(TP+TN)/(TP+FP+TN+FN). out of all customers, the exact predicted values as churned or not churned.

Specificity is calculated using TN/(TN+FP). Exactly predicted true negative value such as retained customers out of all customers.

Sensitivity or Recall is calculated as TP/(TP+FN). exactly predicted true positive values which is churned customers out of all the customers.

ROC curve helps to assess the performance of the model by using graphical illustration. The plot shows the False positive rate and True positive rate.

```
> library(rpart.plot)
> tree <- rpart(Churn ~ ., data = training_set_DT, method = "class")
> tree
n= 5274

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 5274 1402 No (0.73416761 0.26583239)
   2) Contract=One year,Two year 2377  160 No (0.93268826 0.06731174) *
   3) Contract=Month-to-month 2897 1242 No (0.57128064 0.42871936)
     6) InternetService=DSL,No 1317  374 No (0.71602126 0.28397874) *
     7) InternetService=Fiber optic 1580  712 Yes (0.45063291 0.54936709)
      14) tenure>=14.5 842  347 No (0.58788599 0.41211401) *
      15) tenure< 14.5 738  217 Yes (0.29403794 0.70596206) *
>
```

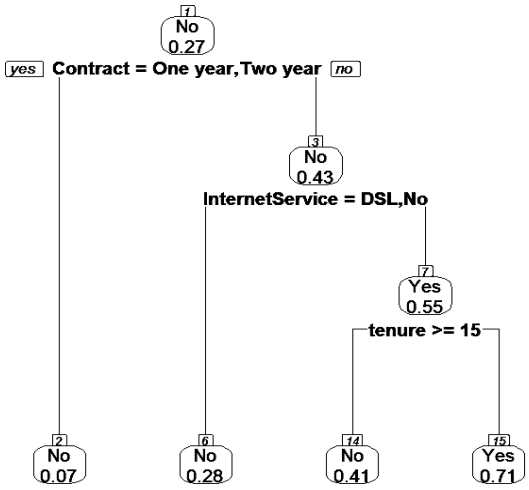


Figure 4: Decision Tree Model with important features

The created decision tree model is depend on those important features such as contract, internet provider and tenure. Furthermore the number of nodes are representing the probability of churned and non-churned customers. In figure 4, The top node has percentage of 27. Since I used a sampling method to get a training set with balanced classes, without using any model there is a 27 % chance for each customer to churn.

As we tend to move down the plot, we will see that chance to churn of customers that have one or 2 the chance of churn is 0.55; otherwise for those whose internet provider is Fiber optic (the last possibility of internet provider) it's0.28.

Customers whose net service supplier don't seem to be Fiber optic and square measure customers of telecommunication a minimum of twelve months have chance to churn is 0.71, otherwise it's 0.41. following result indicates the likelihood of individual client to be churned or not churned.

3.6.2 Estimating Model Accuracy

It is important to consider the accuracy of the particular model before the configuration of test options in a test harness.

3.6.3 Applying k-Fold Cross Validation:

The k-fold cross-validation technique involves breaking the given dataset into k-subsets. for every set is

command out whereas the model is trained on all different subsets. This method is completed till accuracy is set for every instance within the dataset, associate degreed an overall accuracy estimate is provided. It is a sturdy technique for estimating accuracy, and therefore the size of k and tunes the number of bias within the estimate, with in style values set to 3, 5, 7 and 10.

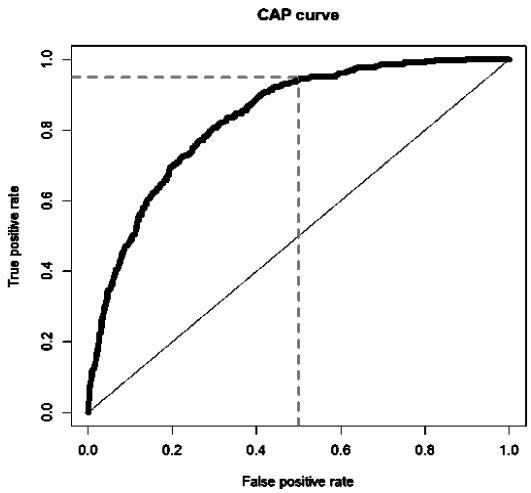
3.7 Performance Evaluation

3.7.1 CAP curve:

The Cumulative Accuracy Profile is able to identify the best model in churn prediction and the prediction of models' capability. The CAP of the model signifies the additive range of positive results on the coordinate axis vs. the respective additive range of a classifying parameter on the x- axis. The CAP is completely different from the Receiver Operator Characteristic (ROC) curves as mythical being curves plot between TP Rate and FP Rate of classifying customers based on Churned and Non-Churned.

In classification model analysis, the CAP curve analysis compares that model with an ideal classification model and a random classification model. It estimates the model by comparison of the trajectory to the right CAP during which the most variety of positive outputs are achieved directly and to the arbitrary CAP during which the positive outputs are equally distributed.

Figure 5: The CAP curve is useful in evaluating a churn prediction model's capability in optimal allocation of resources.



The model can have a CAP between the right CAP and also the arbitrary CAP with a stronger model tending to the right CAP.

The graph demonstrates however taking advantage of information will facilitate telecommunication to spot the customers that are possibly to churn and provides the corporate chance to forestall it. In figure 5, The diagonal line represents the random state of affairs (without employing a model). In the test set 467 customers out of 1758 churned, that close to 26.56 %. If the corporate paid additional attention to every which way chosen 50% of the customers, solely 50% of those 467 would be addressed .

The curve represents state of affairs mistreatment the delivered model of logistical regression. once the corporate pays attention to 50% of the customers with the very best likelihood of churn computed by the model, 86 of those 467 are addressed .To address these 36% of customers which will churn while not the model, telecommunication would have procure approaching another 36% of the customers. therefore just in case of addressing 50% of the customers, the savings from mistreatment of the model represents the value of addressing 36% of the customers. As a matter of truth, the a lot of is that the curve representing the model state of affairs within the high left corner, the higher is that the model.

3.7.2 ROC curve:

ROC Curve is employed to examine how well the classifier will separate positive and negative cases. On the x axis is FP rate, that is that the magnitude relation between the quantity of negative occurrences wrongly classified as FP and therefore the total number of actual negative occurrences.

```
> ##Computing area under the ROC curve
>
> pr <- prediction(prob_pred_LR, test_set$Churn)
> auroc <- performance(pr, measure = 'auc')
> auroc <- auroc@y.values[[1]]
> auc_rounded <- round(auroc,2)
> print(paste('Area under the receiver operating characteristic curve is', auc_rounded))
[1] "Area under the receiver operating characteristic curve is 0.83"
```

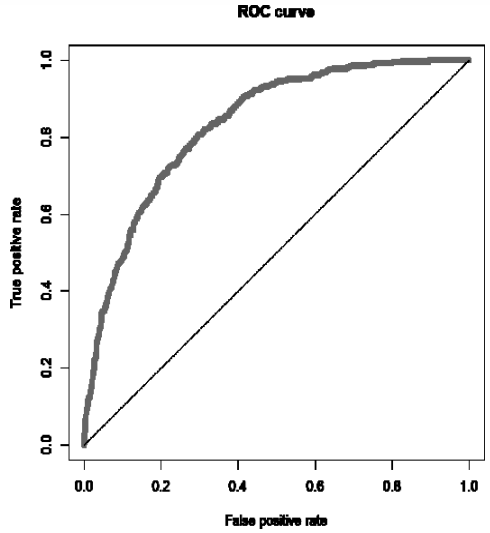


Figure 6: The Area Under Curve (AUC) indicates the models ability to correctly classify those who churned and those who did not.

On the y axis is TP rate, that it the magnitude relation among the number of properly classified positive events and therefore the total number of actual positive events. The obtained ROC is 0.83 for the classification model which is under the category of "Good" shows in table 3.

In figure 6 shows that the diagonal line represents the random model meaning situation while not tested data regarding patterns found within the knowledge.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   No  Yes
##        No  1429  283
##        Yes  119  277
##
##                Accuracy : 0.8093
##                  95% CI : (0.7919, 0.8259)
##     No Information Rate : 0.7343
##     P-Value [Acc > NIR] : 0.0000000000000004604
##
##                   Kappa : 0.4608
##
##  Mcnemar's Test P-Value : 0.0000000000000004304
##
##             Sensitivity : 0.4946
##             Specificity : 0.9231
##          Pos Pred Value : 0.6995
##          Neg Pred Value : 0.8347
##              Prevalence : 0.2657
##          Detection Rate : 0.1314
##    Detection Prevalence : 0.1879
##       Balanced Accuracy : 0.7089
```

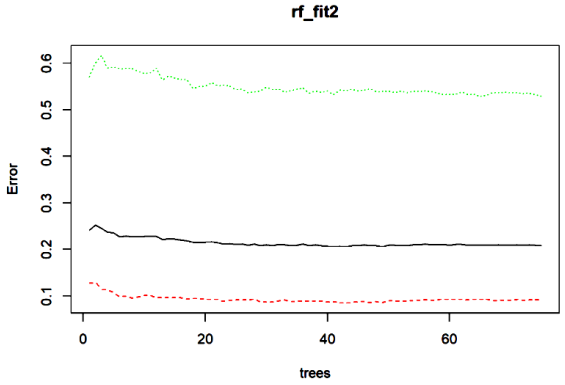


Figure 7: Indicates the performance of the Random forest

fig 7 shows The performance is somewhat similar to the decision tree model. The false negative rate is low (1445 correct vs. 103 incorrect) but the false positive rate is rather high (272 correct vs. 288 incorrect).

The curve represents the situation, within which is that the churn (positive class) foretold by the obtained model (the delivered provision regression).

Table 3: AUROC ranges for the model

AUROC	Category
0.9-1.0	Very good
0.8-0.9	Good
0.7-0.8	Fair
0.6-0.7	Poor
0.5-0.6	Fail

The area under the roc curve (AUROC) ought to be between 0.5 and 1.0. This area could be a measure of the predictive accuracy of a model. an AUROC adequate to 0.5 (i.e. sloping with the diagonal line) which indicates the random classification model. As a matter of reality, this space ought to be larger than 0.5 for a model to be acceptable; a model with AUROC of 0.5 or less is worthless.

4. CONCLUSION AND FUTURE WORK

The importance of this paper in telecom industry is to aid the telecom sector to make profit. Churn Prediction is the key sources of earnings to the industry. The projected LOGISTREE applied statistical model is meant to increase accuracy and also to improve the performance of the created model with the correct dependent options. Fixing the threshold in the acceptable range produce improved accurate results of predicting churn or non-churned customers. Predicting the chance of the customer to churn is taken into account as a necessary facet within the firms. This may stop the customers from going away the corporate, therefore increasing the profit and improving the status of the corporate. As a future scope of the work, client churn will be enforced with totally different techniques to retain the customers.

REFERENCES

1. Hemlata Jain, Ajay Khunteta ,SumitSrivastava Churn Prediction in Telecommunication using Logistic Regression and Logit Boost , International Journal on Computational Intelligence and Data Science , 2020

2. Arifin, S., &Samopa, F. (2018). Analysis of churn rate significantly factors in telecommunication industry using support vector machines method. In Journal of Physics:Conference Series (Vol. 1108, pp. 012018). IOPPublishing.

3. Choudhari, A. S., &Potey, M. (2018). Predictive to prescriptive analysis for customer churn in telecom industry using hybrid data mining techniques. 2018 Fourth international conference on computing communication control and automation (ICCUBEA) (pp. 1–6).IEEE.

4. Rajamohamed, R., &Manokaran, J. (2018). Improved credit card churn prediction based on rough clustering and supervised learning techniques. Cluster Computing, 21(1),65–77.

5. Saran Kumar, A. and Chandrakala, D., 2016. "A Survey on Customer Churn Prediction using Machine Learning Techniques". International Journal of Computer Applications, 975,p.8887.

6. Saghir, M., Bibi, Z., Bashir, S., & Khan, F. H. (2019). Churn prediction using neural network based individual and ensemble models. 2019 16th International Bhurbanconference on applied sciences and technology (IBCAST) (pp.634–639).IEEE.

7. Bi, W., Cai, M., Liu, M. and Li, G., 2016. "A big data clustering algorithm for mitigating the risk of customerchurn".IEEE Transactions on IndustrialInformatics, 12(3),pp.1270-1281.

8. Brandusoiu I, Toderean G, Ha B. Methods for churn prediction in the prepaid mobile telecommunications industry. In: International conference on communications. 2016. p.97–100.

9. Jayaswal, P., Prasad, B.R., Tomar, D. and Agarwal, S., 2016. An Ensemble Approach for Efficient Churn Prediction in Telecom Industry.International Journal of Database Theory and Application, 9(8), pp.211-232.

10. Idris, A. and Khan, A., 2017. "Churn prediction system for telecom using filter–wrapper and ensemble classification". IEEE, 60(3),pp.410-430.

11. Anjum, A., Zeb, A., Afridi, I.U., Shah, P.M., Anjum,A., Raza, B. and Anwar, Z., 2017. "Optimizing Coverage of Churn Prediction in Telecommunication Industry".International Journal of Advanced Computer Science and Applications, 8(5),pp.179-188.

12. Azeem, M. and Usman, M., 2018. A fuzzy based churn prediction and retention model for prepaid customers in telecom industry. International Journal of Computational Intelligence Systems, 11(1),pp.66-

78.
13. Ahmad, A.K., Jafar, A. and Aljoumaa, K., 2019."Customer churn prediction in telecom using machine learning in big data platform". Journal of Big Data, 6(1),p.28.Springer