Telco churn Prediction

MAT394 – MACHINE LEARNING THROUGH R

(Spring 2021)

Bachelor of Technology

In

Electronics and Communication Engineering/Electrical and Electronics Engineering

Submitted by

Datla Gautam Varma (1710110099)

Rohit Kambampati (1710110168)

Ritish Jakkireddy (1810110188)

Under supervision of

Prof Niteesh Sahni

Department of Mathematics

School of Natural Sciences

SHIV NADAR UNIVERSITY

ABSTRACT:

In order to increase the revenue generation base for telco companies, it is key to attract new customers and at the same time avoid churn. Here, when we talk of the word churn, we are discussing about the customers that are going to discontinue from services provided by the company. So, customer churn continues to be one of the most important factors for every communication service provider to consider. Now the question we tend to face next is if we can predict the churn for a telco company from past data? The answer is yes! Data mining and machine learning have made it possible. Machine learning and artificial intelligence are playing a pivotal role in the present-day scenario. It has attained such importance mainly because of its wide range of applications and its incredible ability to adapt and provide solutions to complex problems efficiently, effectively and quickly.

So, in this project first we try to explore the dataset and then using the existing machine learning algorithms such as logistic regression, random forest classifier. decision tree algorithms on the data set we try to figure out which algorithm would turn out to be the best fit for the model giving the least possible errors while predicting. The coding for this project has been done using r in RStudio.

INTRODUCTION:

Customer churn is the loss of clients or customers for a company. Churn is an important business metric for subscription-based services such as telecommunication companies. For such companies to improve their revenue it is important that they predict the churn. Now all the churn data and various factors associated with churn when quantified in the form electronic records, can be used to uncover trends and associations. Machine learning, in particular, can use this data to predict customer churn of a company and identify the most important features among them.

PROBLEM STATEMENT:

Despite the fact that we are familiar with a large number of algorithms, no consensus has emerged to direct the selection of new algorithms for use in churn prediction for telecommunication companies. Although it is possible to select optimal algorithms for research questions and to replicate algorithms in various similar datasets, the understanding and judgement for algorithm implementation is extremely difficult. Now, this is what motivated me to take up this project. In this project I try to build a prediction model that would suit the best by testing out various algorithms on the data set and finally choosing the best fit that would give us minimal error.

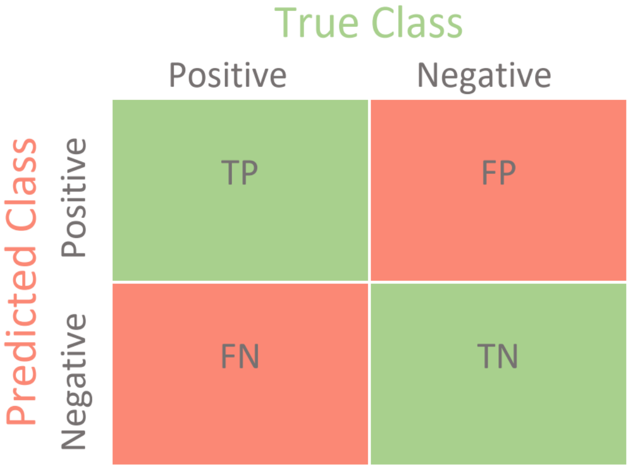
LITERATURE REVIEW

EXPLORATORY DATA ANALYSIS:

It is an approach to analyse the datasets to summarize the data, discover patterns, etc often using data visualisation techniques. For example, using boxplots to identify and remove outliers, visualize and understand the relationship between different features of our data set and the corresponding output.

CONFUSION MATRIX, ROC and AUC:

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. Here we deal with 4 values, tprate, fprate, tnrate, fnrate namely, true-positive rate, true-negative rate, false-positive rate and false-negative rate.



		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Fig1. Structure of a confusion matrix

Now to evaluate the performance of a classification model we use the ROC curve (receiver operating characteristic curve). Using ROC, we calculate the tprate vs fprate at various decision thresholds. Now as an example let's assume that we have obtained the ROC curves for two random ML classifier models. How do we choose the better of the two from ROC?

We use the area under ROC curve. AUC stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve. AUC provides an aggregate measure of performance across all possible classification thresholds. And the main reason to choose AUC is because it is invariant to both scale and classification threshold if any changes occur.

LOGISTIC REGRESSION:

Logistic regression is basically a supervised classification algorithm based on the concept of probability. Logistic regression models the data using the sigmoid function.

$$\mathbf{f(x) = 1/1+e^{-x}}$$

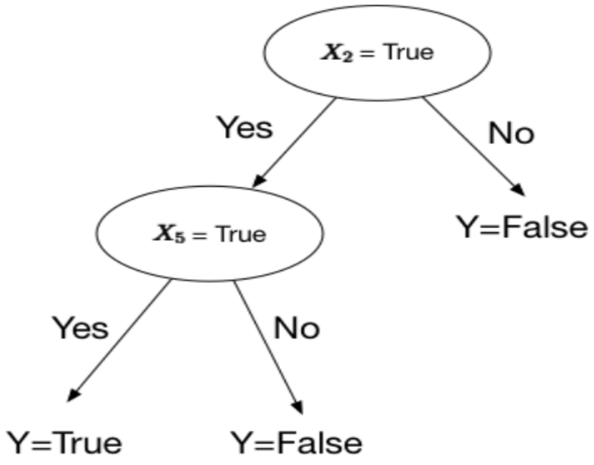
Logistic regression becomes a classification technique only when a decision threshold is brought into the picture. The setting of the threshold value is a very important aspect of Logistic regression and is dependent on the classification problem itself. The decision for the value of the threshold value is majorly affected by the values of precision and recall. Cost function J for logistic regression is,

$$\mathbf{J(\beta) = \Sigma\text{-}y_i \log(h(x_i)) -(1-y_i) \log(1-h(x_i))}$$

Here, h(x) is conditional probability, the probability that given an x_i we get an output yi =1.

DECISION TREE:

Decision Trees are a type of Supervised Machine Learning where the data is continuously split (beginning from decision node) according to a certain parameter. Here, leaves are the final outcomes. If we use decision tree classifier, the decision variable should be categorical.



TELCO CUSTOMER CHURN DATASET

The data was downloaded from IBM Sample Data Sets for customer retention programs. The Telco customer churn dataset contains information about a fictional telco company that provided home phone and Internet services. It provides such services to 7043 customers in California. Multiple important demographics which are included for each customer in the data set are payment method, monthly charges, phones service, internet service, etc.

Each row represents a customer and each column contain customer's attributes, so in total we have 7043 rows (customers) and 21 columns (features) and the "Churn" column is our target.

The data set includes information about:

1) Churn: Customers who left within the last month (1 column)

2) Phone, multiple lines, internet, device protection, online backup, online security, tech support, streaming TV and movies: Services that each customer has signed up for (9 columns)

3) Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges (6 columns)

4) Demographic info about customers – gender, age range, and if they have partners and dependents (5 columns)

METHODOLOGY

I) EDA, Cleaning and pre-processing the data:

To begin with we perform EDA on the dataset. On watching all the features carefully from the dataset, we find categorical features and numerical features. So, to begin with we have used stacked histograms to show the relations of features w.r.t churn. Also, we plotted the density function for all the numerical features. We also plot the output, number of customers left(churn) vs Tenure (Time period) and correlation plot for numerical features as well.

Once we analysed the data properly the next step is to clean and filter out the data set before we split the data into test and train data. So, to begin with we removed the NA values from data. Next, we identify outliers of the data set for all features and remove the outliers. Also, if we look at the data set in the Mutiple lines feature, online security feature we don't have simple yes / no but rather we find no phone service / no internet service instead of simple no for certain samples in the dataset. So, we replace these with "No" so that the entire feature which is categorical has just yes/no I.e., 1/0.

Now, after cleaning the categorical data, now we move onto the binning of the numerical data in dataset. Binning refers to dividing a list of continuous variables into groups and is done to discretize the continuous variables. Bins are easy to analyse and interpret but it leads to loss of some power and information. We should consider distribution of data prior to deciding bin size. Once everything has been done, we finally create a dummy data set and divide the data into test and train data.

ii) Model Building:

We have used logistic regression, decision tree classifier and random forest classifiers in our project. And to choose the best model of the three we use the ROC (receiver operating characteristic curve) and AUC (Area under the curve). After comparing we finally choose the model having the greatest area under the ROC curve.

RESULTS AND CONCLUSIONS

1. Categorical features and their relation w.r.t churn:

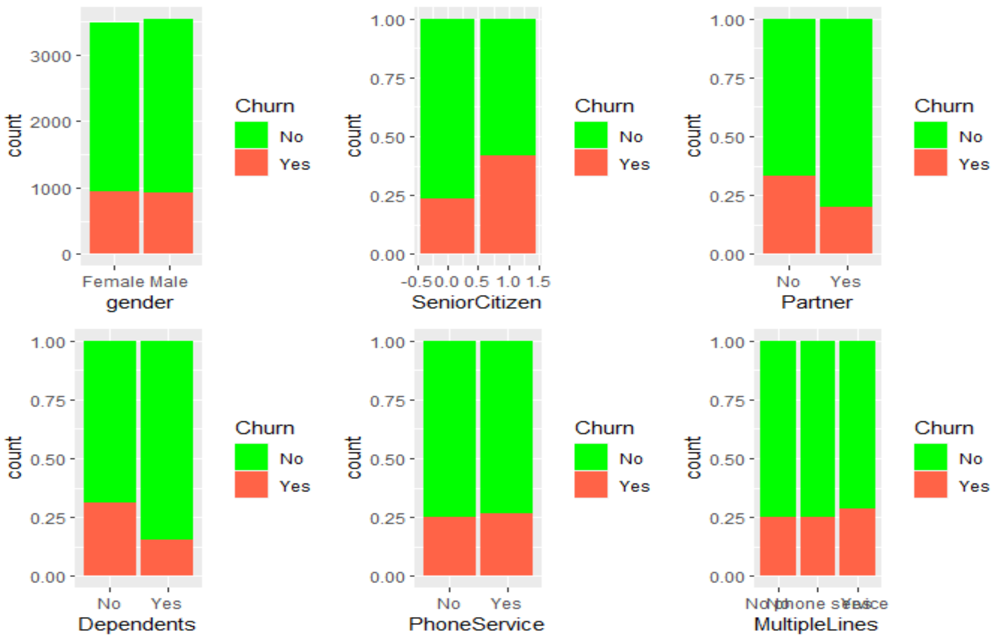


Fig3. Gender, senior citizens, partners, dependents, Phone service and multiple lines w.r.t churn

Gender: The churn percent is almost equal in case of Male and Females

Senior Citizen: Senior citizens have higher churn rate

Partners and dependents: Customers with Partners and Dependents have lower churn rate

Phone Service and Multiple Lines: We can notice from the stacked histograms that we have almost same amount of churn for almost all the cases of phone service and multiple lines.

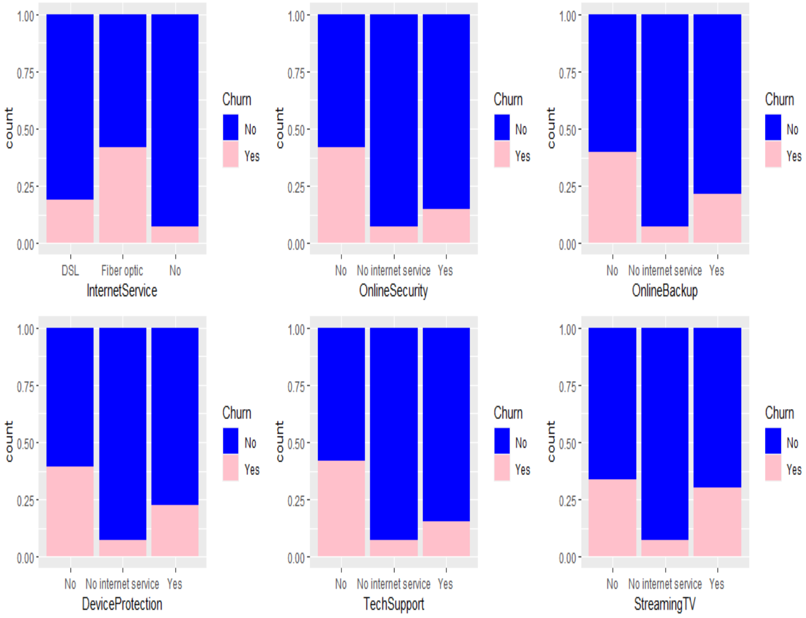


Fig4. Internet service, online security, Online backup, Device Protection, Tech support and Streaming TV w.r.t churn

Churn rate is higher in case of Internet Services using fibre optic when compared to other cases of internet service. Also, Customers having no online security, online backup and Tech Support also come under churned customers.

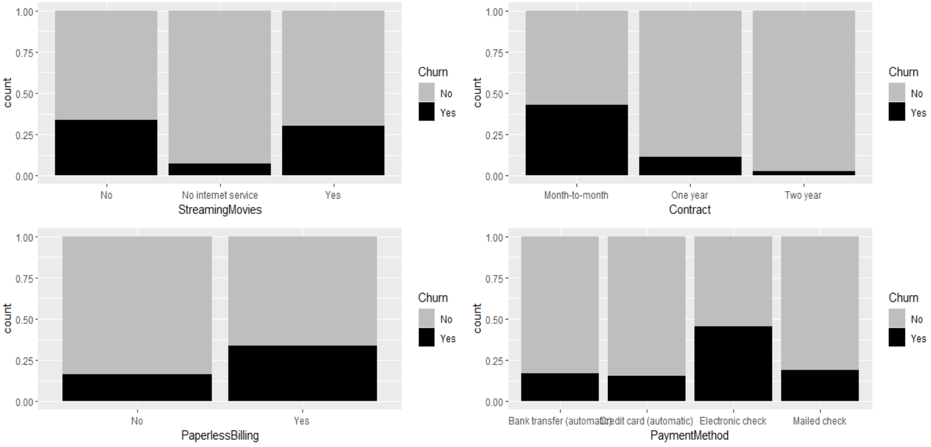


Fig5.Streaming movies, contract, paperless billing and payment method w.r.t churn

Customers with yearly subscription have less churn when compared to people with monthly subscriptions. Churn percent is higher in case of customers having paperless billing option. Also, people with E-check as payment method have more churn than people with other payment methods.

2) Numerical features and their relation w.r.t churn:

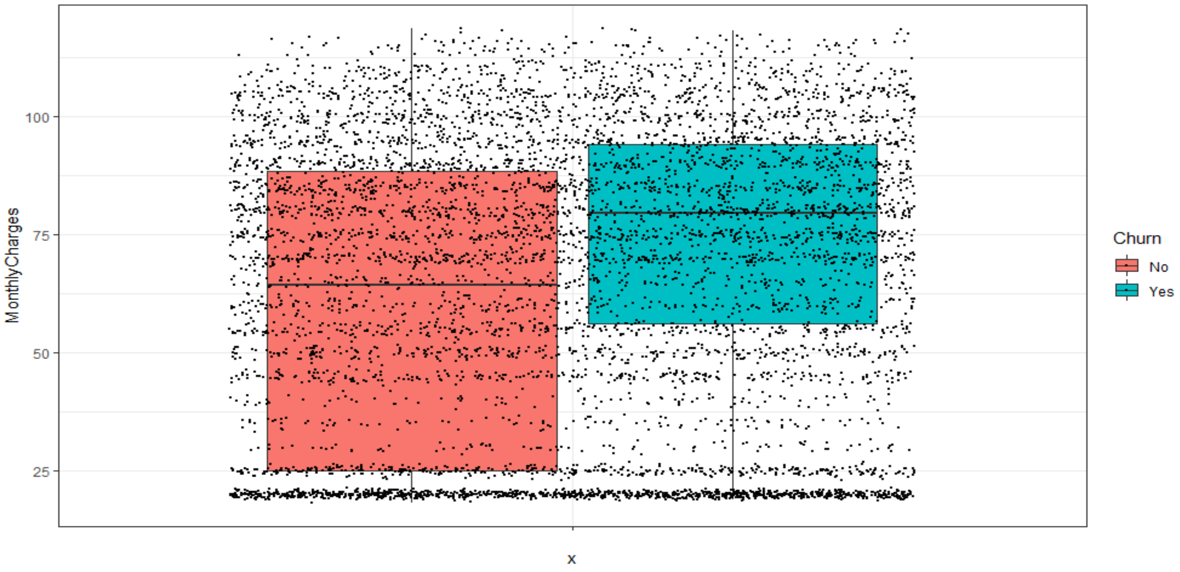


Fig6.Relationship of monthly charges w.r.t churn

Monthly Charges: The median of the monthly charges of churned customers is around 75 and median of customer who haven't left s around 60

Also, we have plotted the sample points so that we can actually visualize the distribution of the box plot.

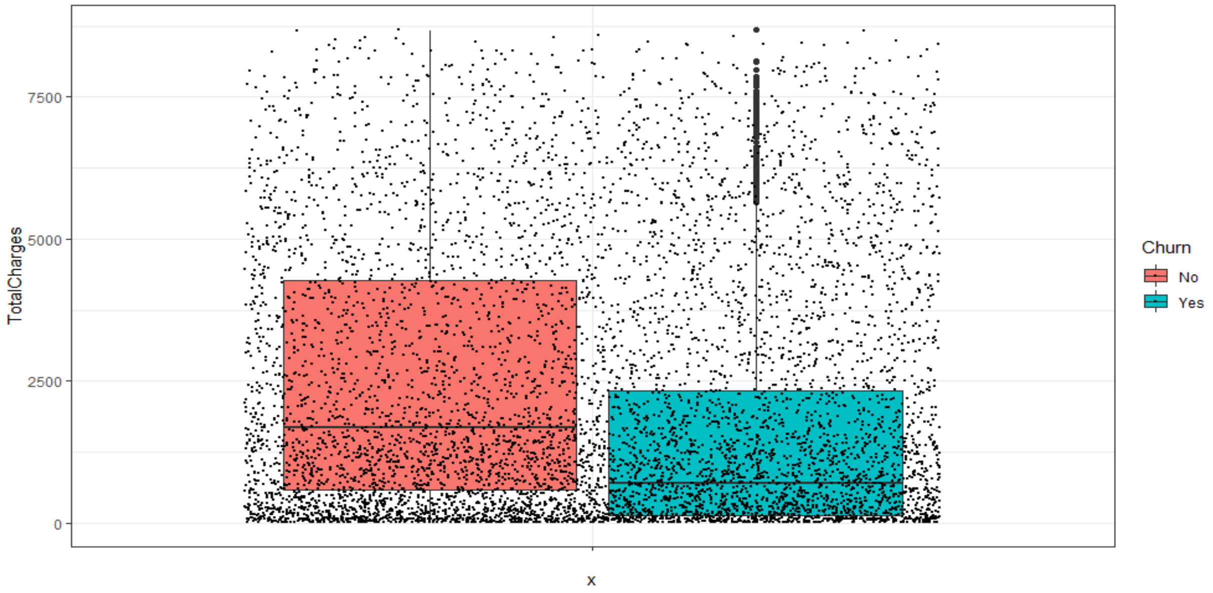


Fig7.Relationship of total charges w.r.t churn

Total Charges: The median of the total charges of churned customers is around 1500 and median of customer who haven't left s around 600.

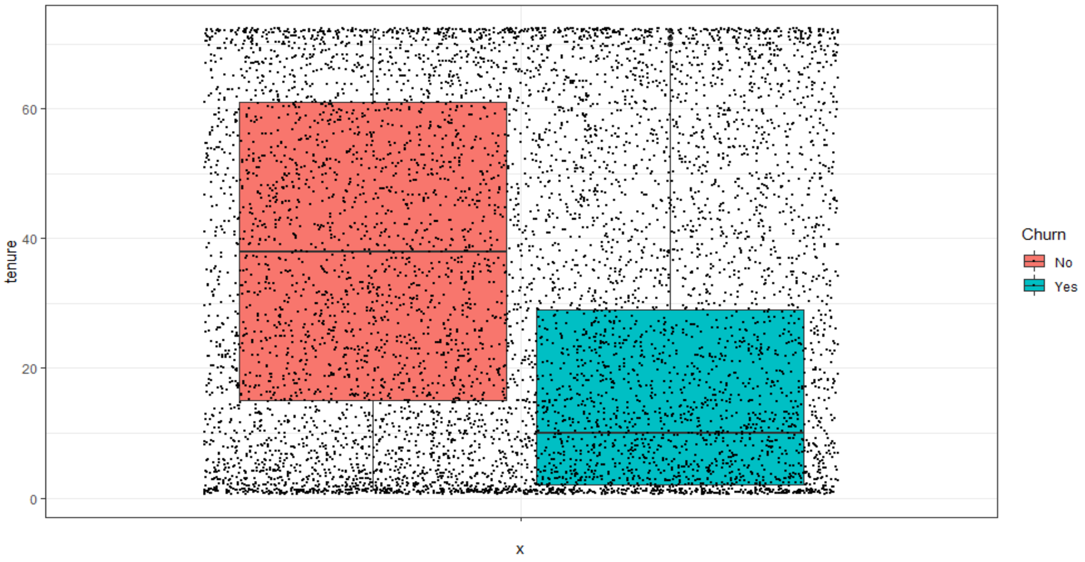


Fig8. Tenure w.r.t churn

Tenure: The median tenure for churned customers is around 10 months and customers who haven't left is around 40 months.

3) Density plots: *Dashed lines here represent the mean of the distribution*

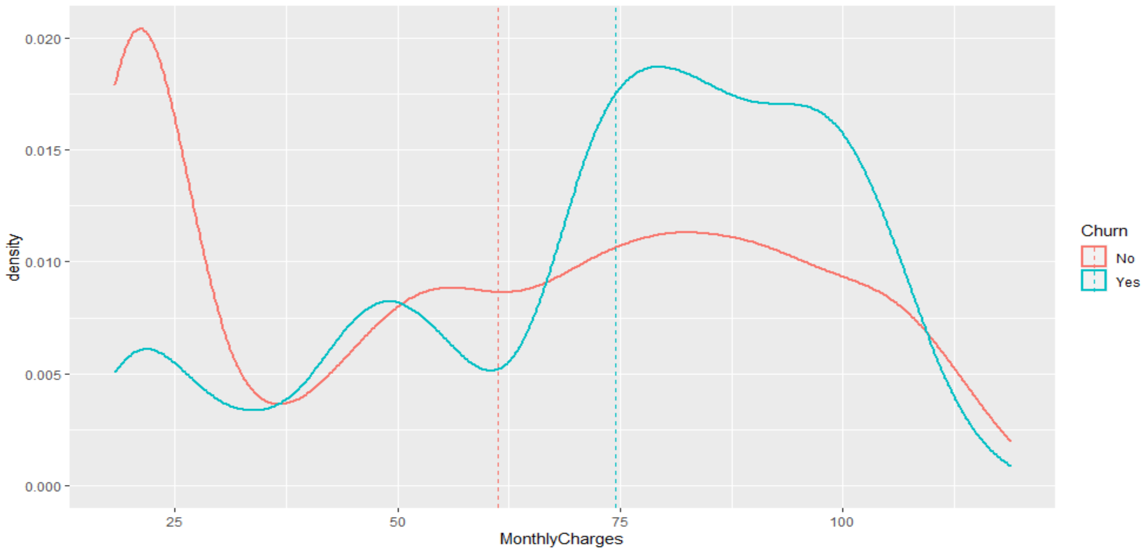


Fig9. Monthly charges density plot

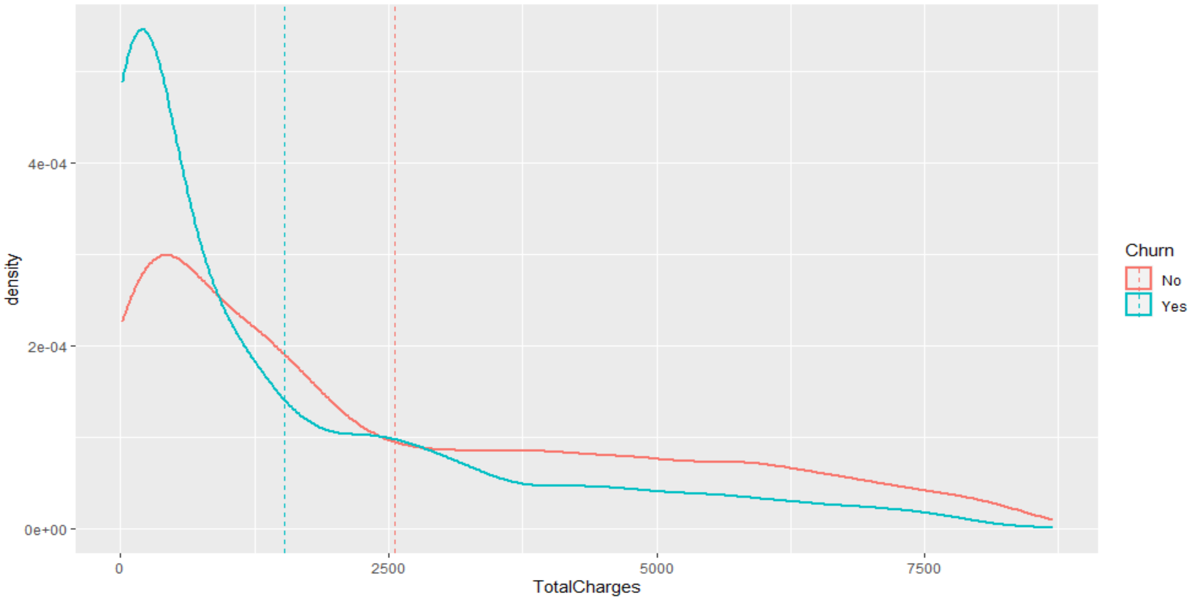


Fig10. Total charges density plot

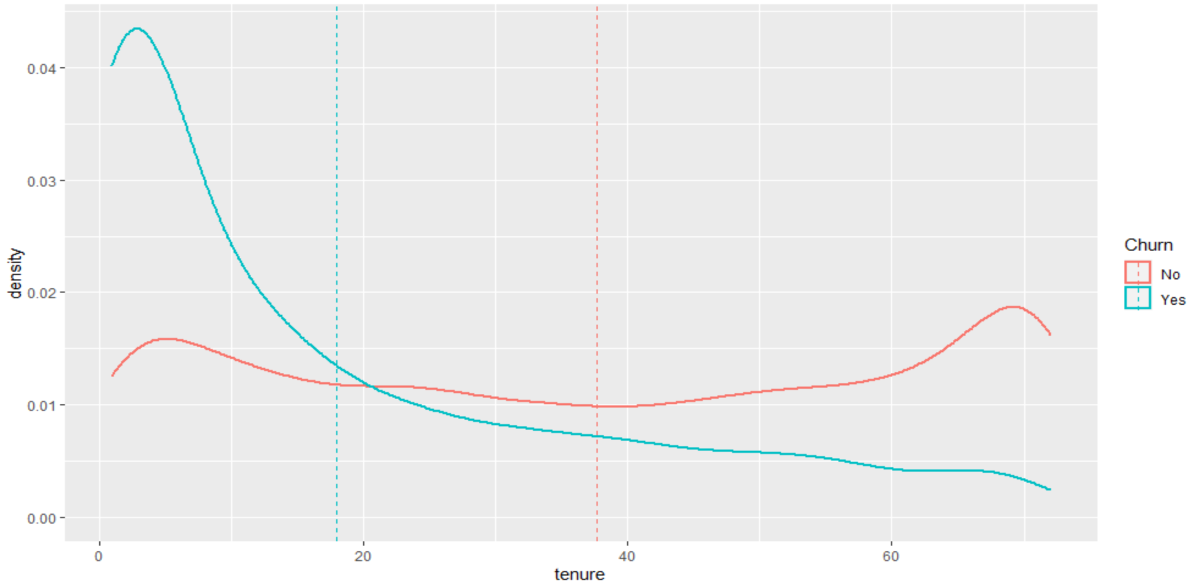


Fig11. Tenure density plot

4) Correlation and outliers:

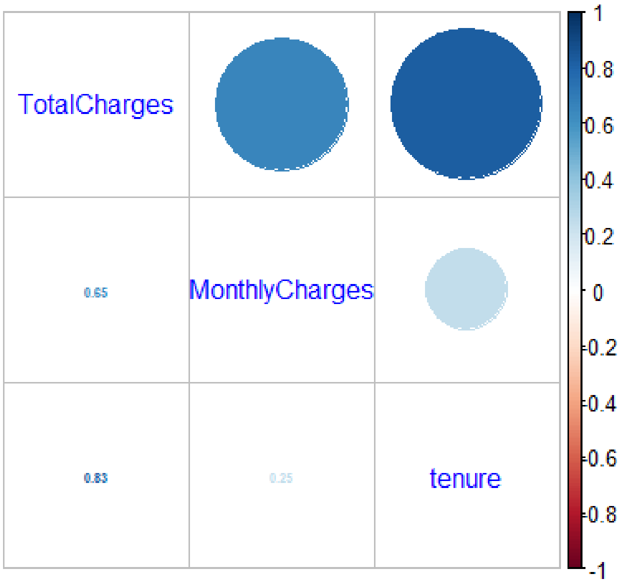


Fig12. Correlation plot

The plot shows high correlations between Total charges & tenure and between Total Charges & Monthly Charges.

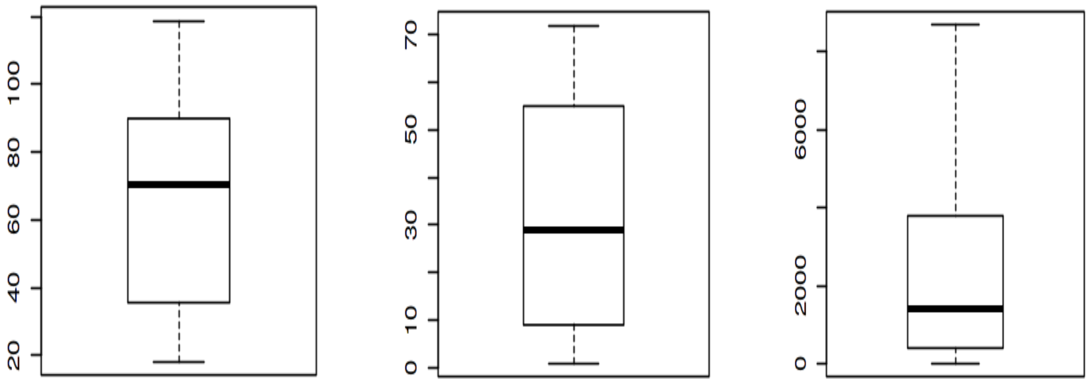


Fig12. Boxplots of Monthly charges, total charges and tenure (with zero outliers)

Looking at the boxplot for all the three numerical features we can clearly see that there are no outliers to be removed from the data set.

If we look at the data set in the Mutiple lines feature, online security feature we don't have simple yes / no but rather we find no phone service / no internet service instead of simple no for certain samples in the dataset. So, we replace these with "No" and clean these features.

Next, we remove NA values from the data set. We can see that all the NA values are from the Total Charges feature.

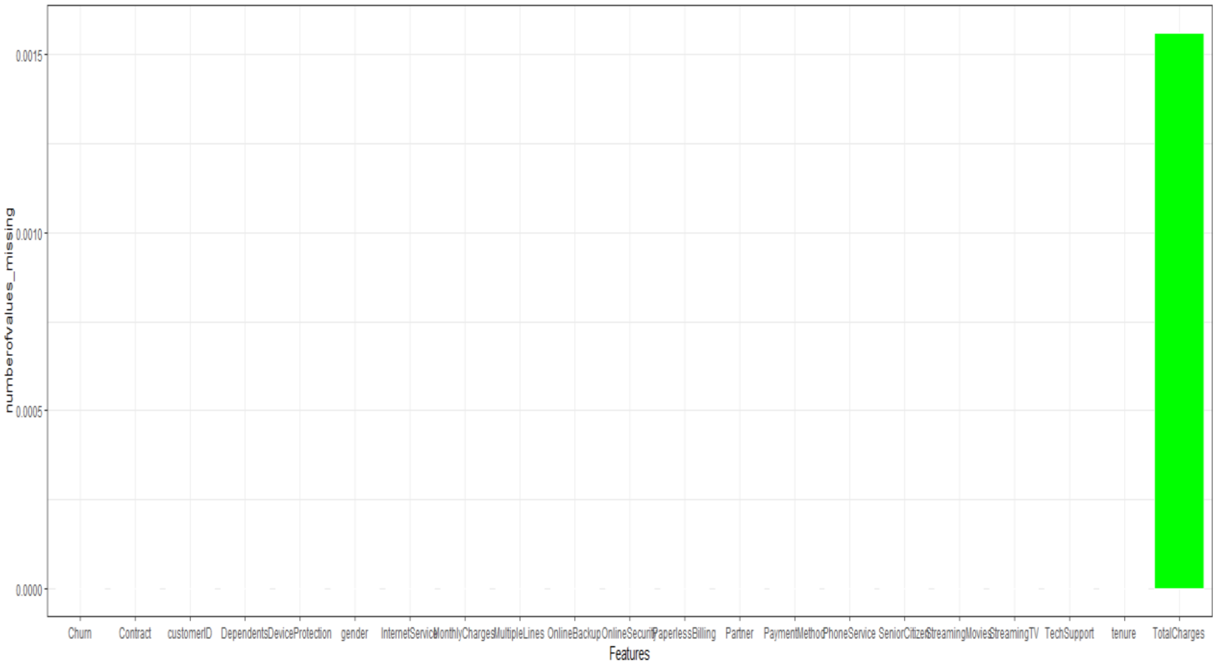
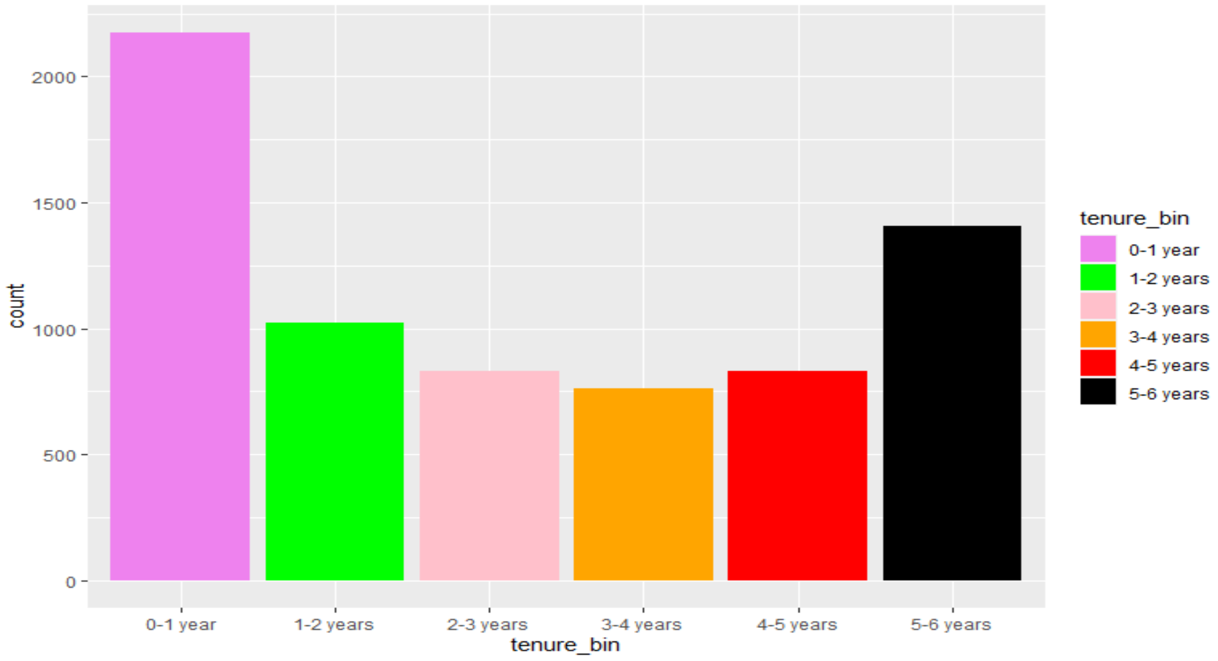


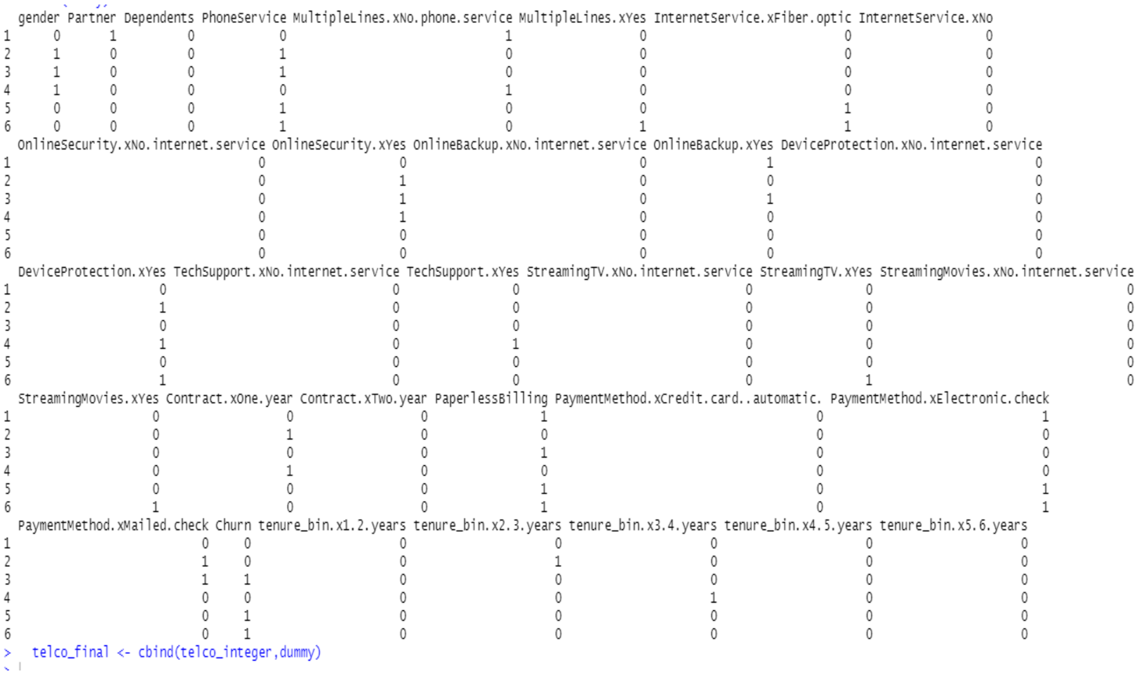
Fig13. Plot showing the number of NA values in each feature

Here we know that the tenure we are talking of represents time period in months. To find better churn patterns w.r.t time we bin the data with 5 levels, with each level represents a bin of tenure in years.



Plot 14. Tenure Bins

Next, we create a dummy table in which all categorical features are converted to integers. And then we finally combine the dummy data frame and the numerical data frame to obtain the final data frame which we will be splitting into test and train data.



```
  gender Partner Dependents PhoneService MultipleLines.xNo.phone.service MultipleLines.xYes InternetService.xFiber.optic InternetService.xNo
1      0       1          0            0                              1                  0                            0                   0
2      1       0          0            1                              0                  0                            0                   0
3      1       0          0            1                              0                  0                            0                   0
4      1       0          0            0                              1                  0                            0                   0
5      0       0          0            1                              0                  0                            1                   0
6      0       0          0            1                              0                  1                            1                   0
  OnlineSecurity.xNo.internet.service OnlineSecurity.xYes OnlineBackup.xNo.internet.service OnlineBackup.xYes DeviceProtection.xNo.internet.service
1                                   0                   0                                 0                 1                                    0
2                                   0                   1                                 0                 0                                    0
3                                   0                   1                                 0                 1                                    0
4                                   0                   1                                 0                 0                                    0
5                                   0                   0                                 0                 0                                    0
6                                   0                   0                                 0                 0                                    0
  DeviceProtection.xYes TechSupport.xNo.internet.service TechSupport.xYes StreamingTV.xNo.internet.service StreamingTV.xYes StreamingMovies.xNo.internet.service
1                     0                                0                0                                0                0                                    0
2                     1                                0                0                                0                0                                    0
3                     0                                0                0                                0                0                                    0
4                     1                                0                1                                0                0                                    0
5                     0                                0                0                                0                0                                    0
6                     1                                0                0                                0                1                                    0
  StreamingMovies.xYes Contract.xOne.year Contract.xTwo.year PaperlessBilling PaymentMethod.xCredit.card..automatic. PaymentMethod.xElectronic.check
1                    0                  0                  0                1                                      0                               1
2                    0                  1                  0                0                                      0                               0
3                    0                  0                  0                1                                      0                               0
4                    0                  1                  0                0                                      0                               0
5                    0                  0                  0                1                                      0                               1
6                    1                  0                  0                1                                      0                               1
  PaymentMethod.xMailed.check Churn tenure_bin.x1.2.years tenure_bin.x2.3.years tenure_bin.x3.4.years tenure_bin.x4.5.years tenure_bin.x5.6.years
1                           0     0                     0                     0                     0                     0                     0
2                           1     0                     0                     1                     0                     0                     0
3                           1     1                     0                     0                     0                     0                     0
4                           0     0                     0                     0                     1                     0                     0
5                           0     1                     0                     0                     0                     0                     0
6                           0     1                     0                     0                     0                     0                     0
>  telco_final <- cbind(telco_integer,dummy)
```

Plot 15. Top 5 rows I.e., head of dummy data frame

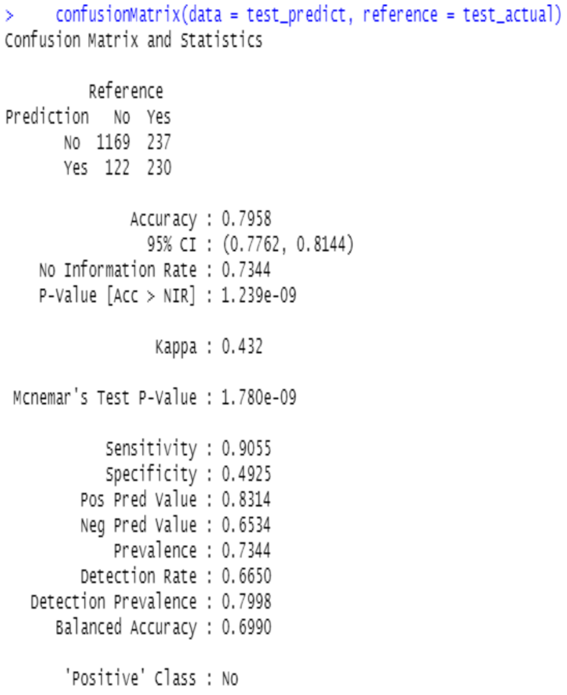
5) Model building:

a. Logistic regression:

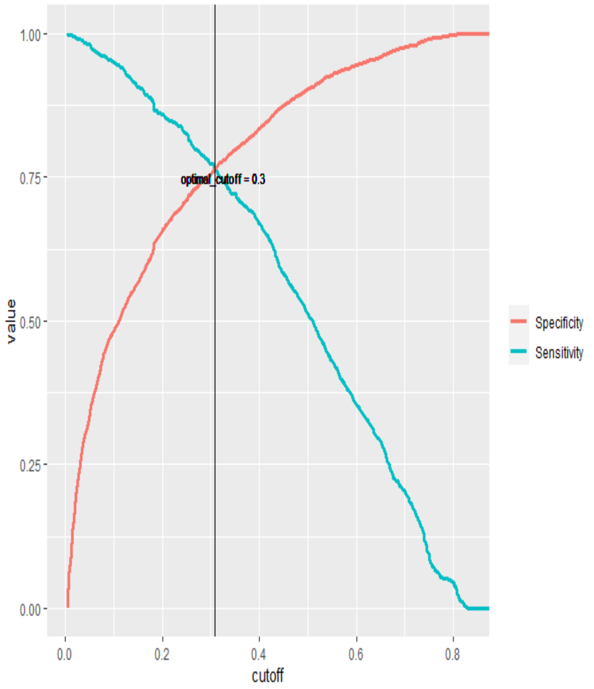
First step is that we use AIC for variable selection. AIC is an iterative process of adding or removing variables, in order to get the best performing model using only a specific subset of features. Next is the VIF (variation inflation factor). Which is used to measure the multicollinearity between predictor variables in a model. Higher the VIF, greater is the correlation of the predictor variable w.r.t other predictor variables. But we always, need to see the significance of the Predictor variable before removing it from our model.

For implementing the logistic regression, we used the glm() function. The Generalized linear models are fit using the glm( ) function. glm(formula, family=family type (link=link function), data=) If the link is binary, we'll be using logit/logistic regression.

Now, we need to find the optimal probability cut-off which will give maximum accuracy, sensitivity and specificity.

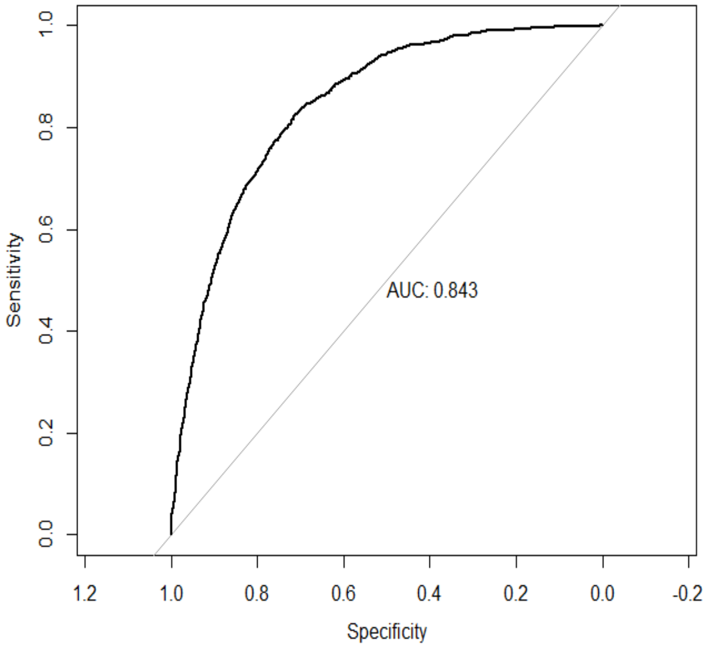


```
>    confusionMatrix(data = test_predict, reference = test_actual)
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1169  237
       Yes  122  230

               Accuracy : 0.7958
                 95% CI : (0.7762, 0.8144)
    No Information Rate : 0.7344
    P-Value [Acc > NIR] : 1.239e-09

                  Kappa : 0.432

 Mcnemar's Test P-Value : 1.780e-09

            Sensitivity : 0.9055
            Specificity : 0.4925
         Pos Pred Value : 0.8314
         Neg Pred Value : 0.6534
             Prevalence : 0.7344
         Detection Rate : 0.6650
   Detection Prevalence : 0.7998
      Balanced Accuracy : 0.6990

       'Positive' Class : No
```



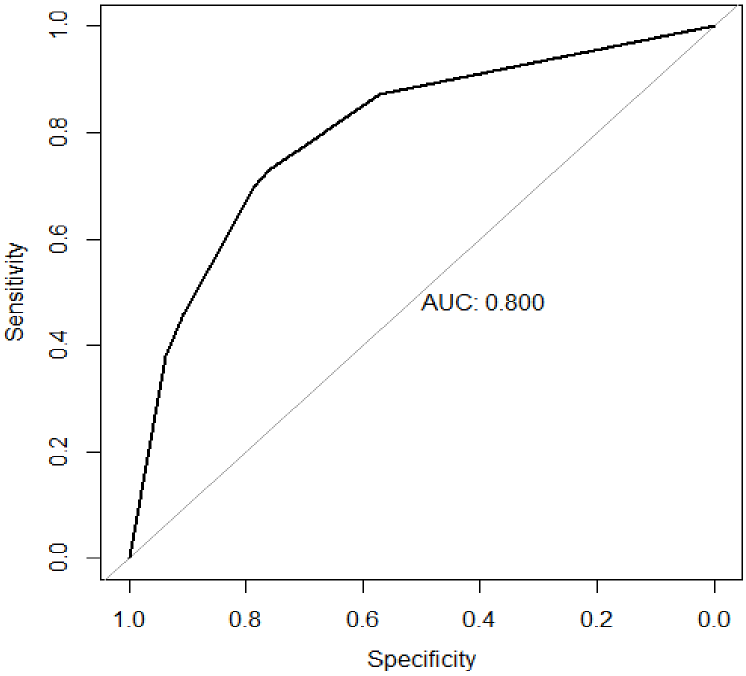
Plot 16. Confusion matrix and statistics for Logistic regression

Plot 17. Selection optimal cut-off threshold in logistic regression

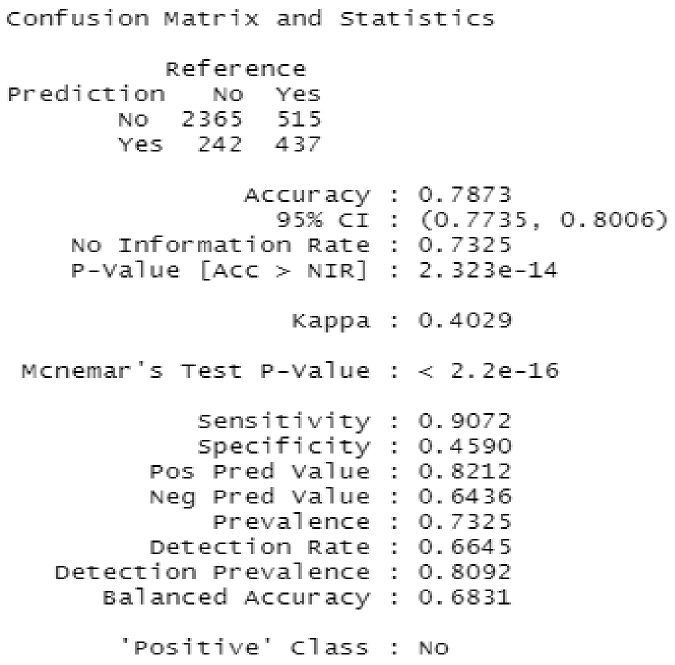


Plot 18. ROC curve for Logistic regression

b. Decision Tree classifier:



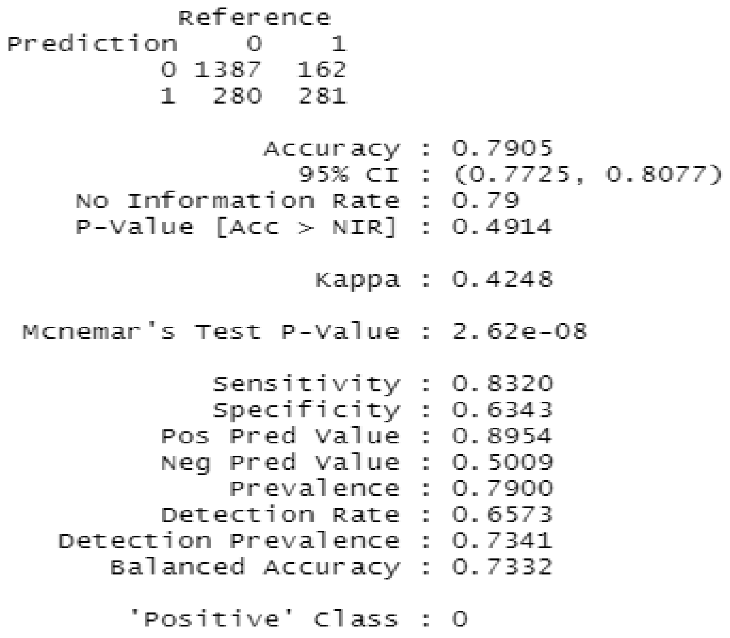
Plot19. ROC curve for decision trees



```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  2365  515
       Yes  242  437

               Accuracy : 0.7873
                 95% CI : (0.7735, 0.8006)
    No Information Rate : 0.7325
    P-Value [Acc > NIR] : 2.323e-14

                  Kappa : 0.4029

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9072
            Specificity : 0.4590
         Pos Pred Value : 0.8212
         Neg Pred Value : 0.6436
             Prevalence : 0.7325
         Detection Rate : 0.6645
   Detection Prevalence : 0.8092
      Balanced Accuracy : 0.6831

       'Positive' Class : No
```

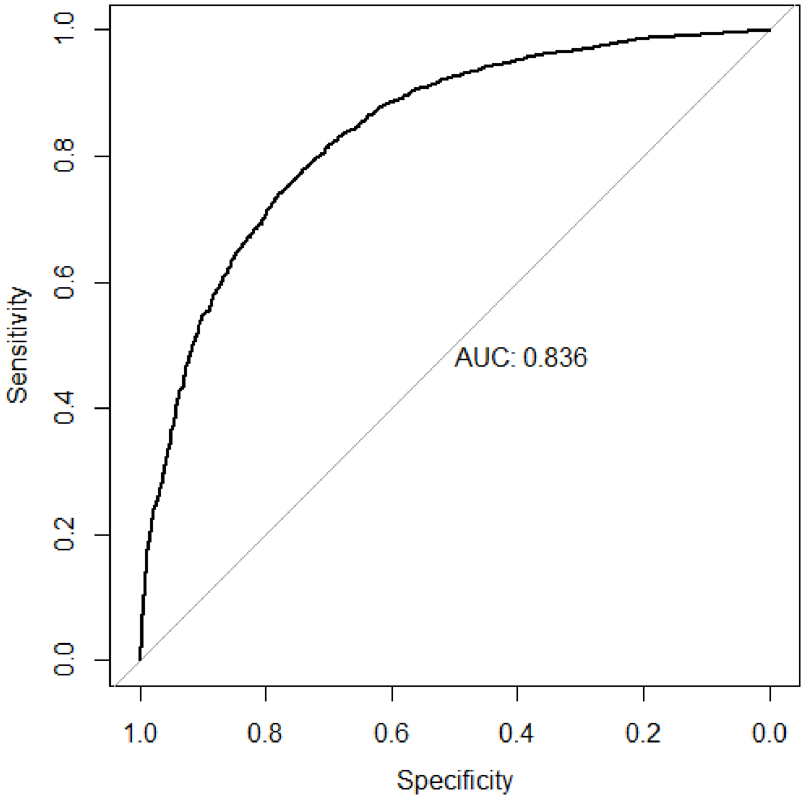
Plot 20. Confusion matrix and statistics for decision trees

C. Random Forest:



```
          Reference
Prediction    0    1
         0 1387  162
         1  280  281

               Accuracy : 0.7905
                 95% CI : (0.7725, 0.8077)
    No Information Rate : 0.79
    P-Value [Acc > NIR] : 0.4914

                  Kappa : 0.4248

 Mcnemar's Test P-Value : 2.62e-08

            Sensitivity : 0.8320
            Specificity : 0.6343
         Pos Pred Value : 0.8954
         Neg Pred Value : 0.5009
             Prevalence : 0.7900
         Detection Rate : 0.6573
   Detection Prevalence : 0.7341
      Balanced Accuracy : 0.7332

       'Positive' Class : 0
```

Plot 21. Confusion matrix and statistics for Random Forest classifier



Plot22. ROC curve for random forest classifier

Model	tprate	fprate	AUC (Area under the ROC)	Accuracy
Logistic Regression	0.653	0.168	0.843	79.58%
Extra Tree	0.643	0.1788	0.800	78.73%
Random Forest	0.501	0.104	0.836	79.05%

Table1. Observations from Plots and Statistics for all three models

Looking at table of observations from the obtained plots and statistics from all the three models, we can clearly figure out that Logistic regression is our choice of best performing model mainly because it has highest accuracy (79.58%), highest tprate of all three (closest to 1), a good fprate (close to zero) and highest Area under ROC (Receiver optimum characteristics curve)

REFERENCES

[1] T. Haifley, "Linear logistic regression: an introduction," IEEE International Integrated Reliability Workshop Final Report, 2002., 2002, pp. 184-187, doi: 10.1109/IRWS.2002.1194264.

[2] S. V. Patel and V. N. Jokhakar, "A random forest-based machine learning approach for mild steel defect diagnosis," 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2016, pp. 1-8, doi: 10.1109/ICCIC.2016.7919549.

[3] A. Navada, A. N. Ansari, S. Patil and B. A. Sonkamble, "Overview of use of decision tree algorithms in machine learning," *2011 IEEE Control and System Graduate Research Colloquium*, 2011, pp. 37-42, doi: 10.1109/ICSGRC.2011.5991826.

[4] BlastChar. (2018, February 23). *Telco Customer Churn*. Kaggle. https://www.kaggle.com/blastchar/telco-customer-churn.

[5] Secunsexto. (2018, September 19). *Clustering, churn and Correlation analysis*. Kaggle. https://www.kaggle.com/secunsexto/clustering-churn-and-correlation-analysis.

[6] *Cognos Analytics*. Cognos Analytics - IBM Business Analytics Community. (n.d.). https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113.