# GROUP 9 PROJECT REPORT

CUSTOMER CHURN PREDICTION

Presented by:
Tejaswini Yeruva, Karthik Badiganti, Nikhil Reddy Addula, Venu Dodda

BUSINESS ANALYTICS
MIS 64036

## Abstract

Customer churn is a problem for telecom companies since it is more difficult to find new customers than it is to retain on their existing customers. Companies regularly utilize data mining tools to assist them identify customers who are likely to leave. Based on the historical data, we were able to identify clients who were most likely to churn and provide ABC Wireless Inc with significant help to retain the customers.

## Introduction

Finding the right balance between customer acquisition and retention, two important factors that have a direct effect on a company's financial performance, can be difficult. Because of churn rate affects the company's business, customer retention is probably going to be a key factor.

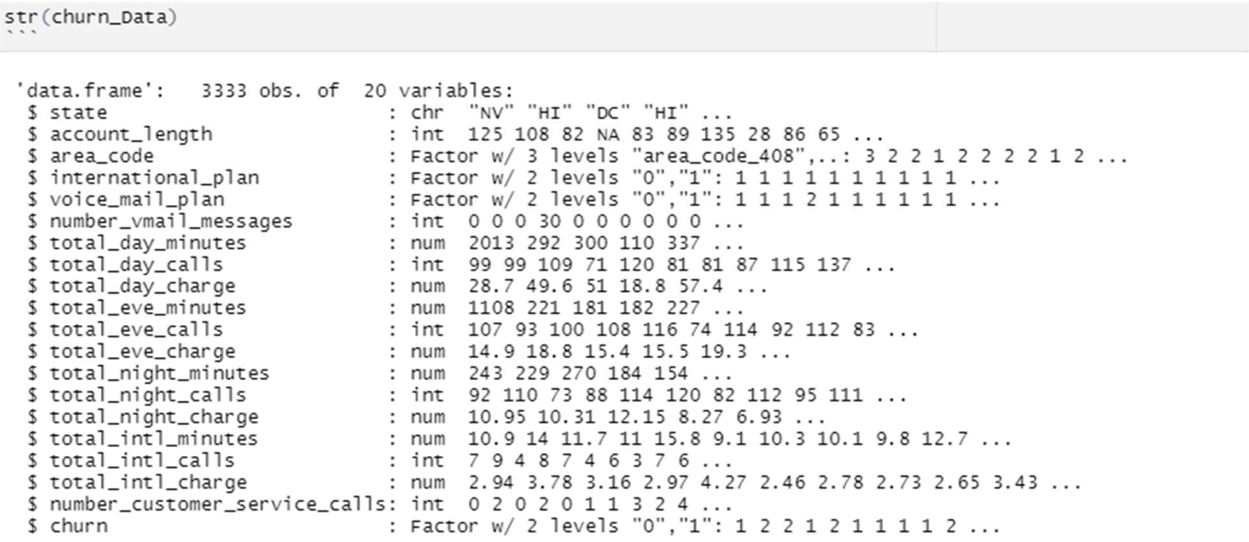
Customers may switch service providers for several reasons, including poor network connections, poor customer support, and expensive monthly plans. These problems can be resolved, among other things, by giving deals or better customer service, so that customers don't leave their company to another service provider.

When we have the analytical skills to evaluate and analyse complicated data and extract relevant information in the modern day, we can use this knowledge to draw patterns and foresee future outcomes.

The goal of this project is to use a predictive model to analyse data, identify trends, and anticipate when a regular client would switch service providers. We used a range of prediction models, including regression, ANOVA, statistics (sensitivity, accuracy), pruning, and plots, to execute our analysis. In this instance, we'll are using a decision tree and logistic regression to build our model.

## Overview

ABC wireless company has provided the following data from which we can infer that There are 3333 observations with around 20 variables.

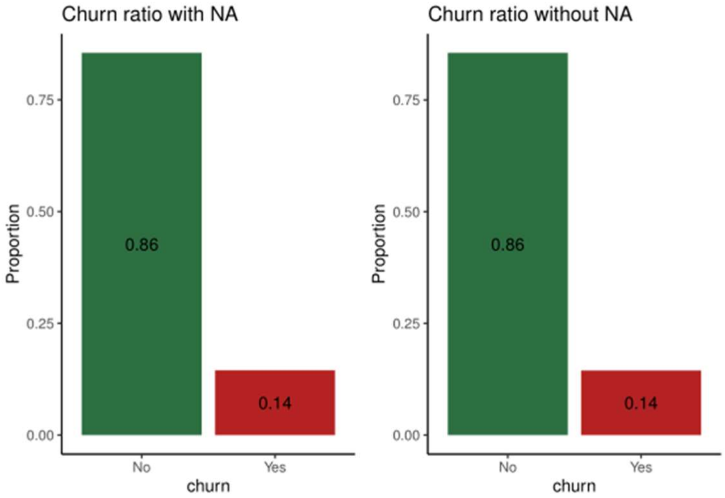


~~~
str(churn_Data)
```
~~~

```
 'data.frame':   3333 obs. of  20 variables:
  $ state                        : chr  "NV" "HI" "DC" "HI" ...
  $ account_length               : int  125 108 82 NA 83 89 135 28 86 65 ...
  $ area_code                    : Factor w/ 3 levels "area_code_408",..: 3 2 2 1 2 2 2 2 1 2 ...
  $ international_plan           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
  $ voice_mail_plan              : Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 1 1 1 ...
  $ number_vmail_messages        : int  0 0 0 30 0 0 0 0 0 0 ...
  $ total_day_minutes            : num  2013 292 300 110 337 ...
  $ total_day_calls              : int  99 99 109 71 120 81 81 87 115 137 ...
  $ total_day_charge             : num  28.7 49.6 51 18.8 57.4 ...
  $ total_eve_minutes            : num  1108 221 181 182 227 ...
  $ total_eve_calls              : int  107 93 100 108 116 74 114 92 112 83 ...
  $ total_eve_charge             : num  14.9 18.8 15.4 15.5 19.3 ...
  $ total_night_minutes          : num  243 229 270 184 154 ...
  $ total_night_calls            : int  92 110 73 88 114 120 82 112 95 111 ...
  $ total_night_charge           : num  10.95 10.31 12.15 8.27 6.93 ...
  $ total_intl_minutes           : num  10.9 14 11.7 11 15.8 9.1 10.3 10.1 9.8 12.7 ...
  $ total_intl_calls             : int  7 9 4 8 7 4 6 3 7 6 ...
  $ total_intl_charge            : num  2.94 3.78 3.16 2.97 4.27 2.46 2.78 2.73 2.65 3.43 ...
  $ number_customer_service_calls: int  0 2 0 2 0 1 1 3 2 4 ...
  $ churn                        : Factor w/ 2 levels "0","1": 1 2 2 1 2 1 1 1 1 2 ...
```

## Data Cleaning & Partition:

We can observe that there are negative values in account length column, assuming that that they might be mistakenly entered negative, hence taking their absolute values.



From the above plot, churn ratio remains the same even after removing NA values. Therefore, removing NA values from the data.

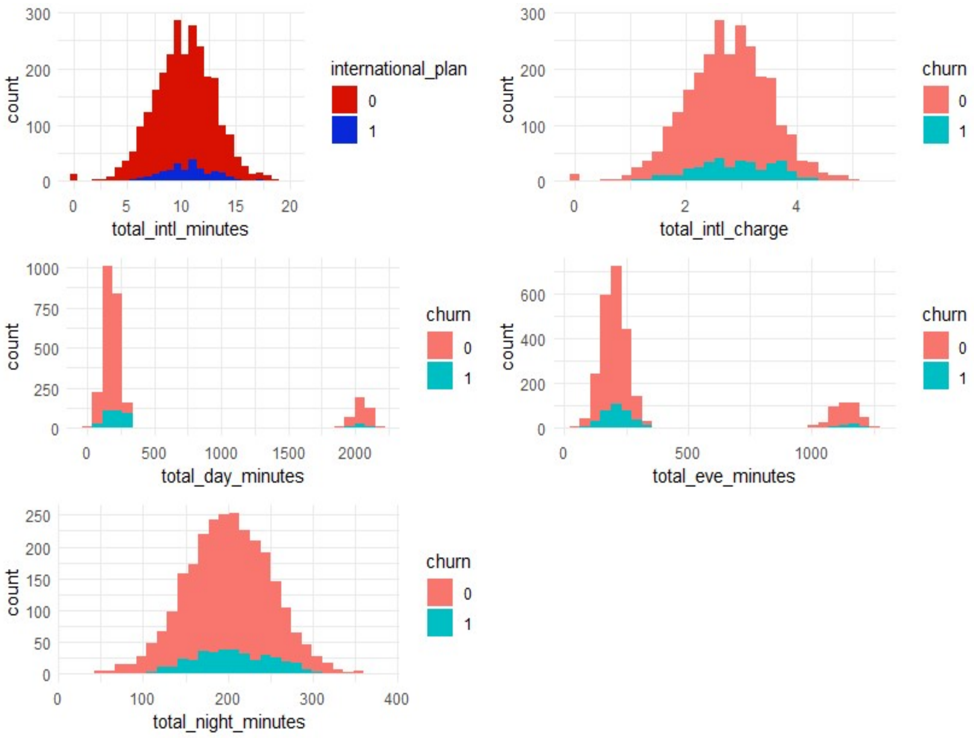
## Data Exploration:

The above data can be explored and analyzed by dividing them into two categories,

1. Customer usage
2. Network Coverage & Customer Service

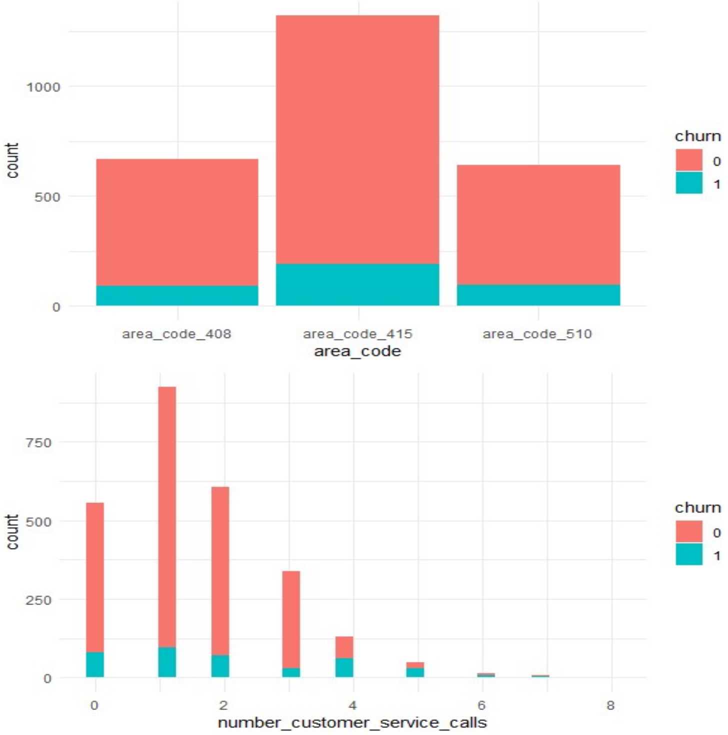
### Customer Usage:

It can be observed from below plots is that there are lot of customers using international calls without international plan and when it comes to churn, there was an increase in churn ratio with increase of usage in international calls. Also, there was a similarity observed between day minutes and evening minutes. Also, the churn customer ratio increases with increase in the usage of night calling.

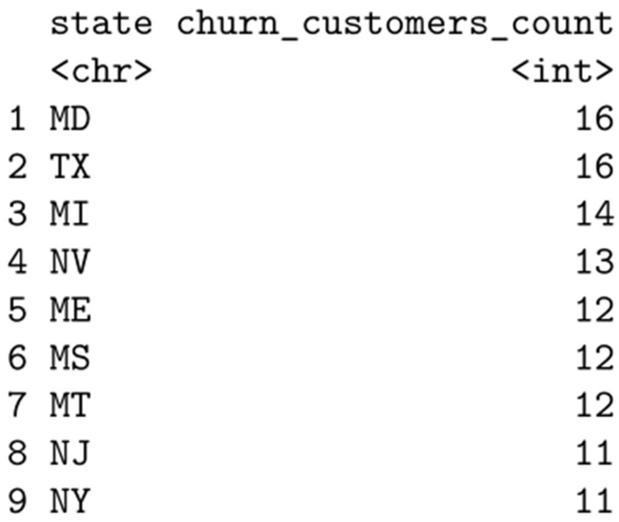


### Network Coverage & Customer Service:

For any network to succeed, there needs to be good network service and better customer service. From the data, it can be observed that area 415 has large number of users and relatively high churn ratio. Also, customers who are having a greater number of customer service calls are also leaving the network. All these can be observed in the below plots.



It can be also observed that states like Maryland and Texas have large number of churn customers.



```
  state churn_customers_count
  <chr>                 <int>
1 MD                       16
2 TX                       16
3 MI                       14
4 NV                       13
5 ME                       12
6 MS                       12
7 MT                       12
8 NJ                       11
9 NY                       11
```

# Model Building

## Strategy

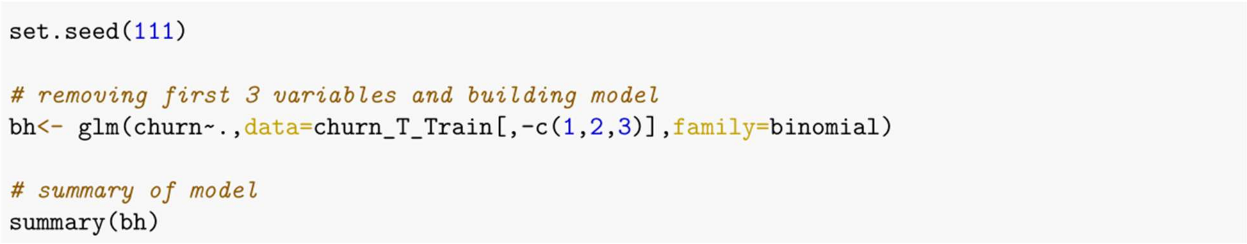
Multimodal classification is the technique of separating a sample into two categories using a classifier. Since the attribute in this data is categorical and the output for this model is a greater chance or likelihood of odds between 0 and 1, we will be using both logistic regression and decision trees to address this issue by comparing the performance matrices of both models to see which is more efficient.

## Data Partition:

Before building the model, the data is divided into two parts, 70% of train data se and 30% of validation data set.

It is a classification model; we are developing two models and then compare their Area under curve to choose the best model.

## Logistic Regression:



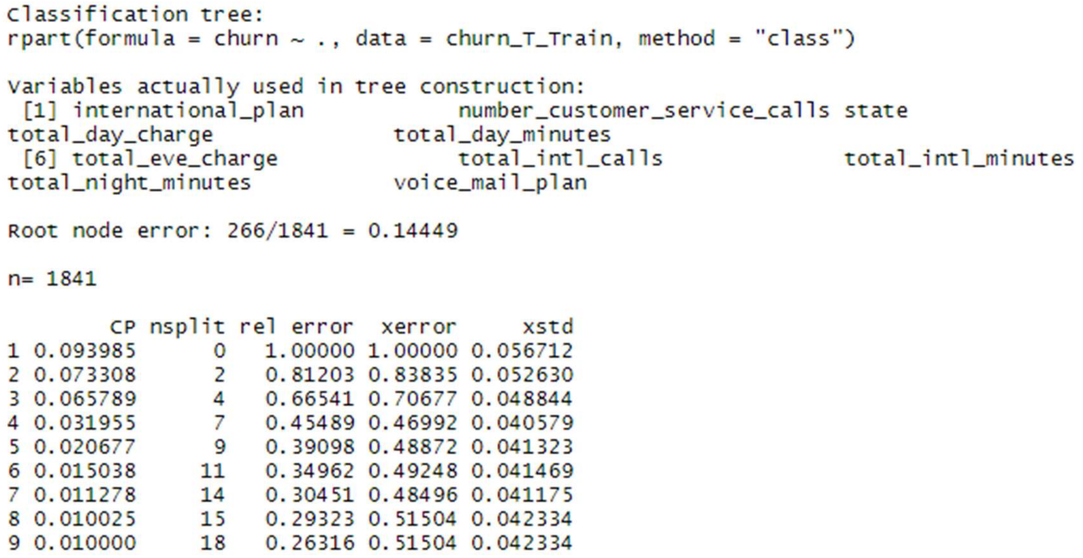
```
set.seed(111)

# removing first 3 variables and building model
bh<- glm(churn~.,data=churn_T_Train[,-c(1,2,3)],family=binomial)

# summary of model
summary(bh)
```

Reasons for using Logistic Regression:

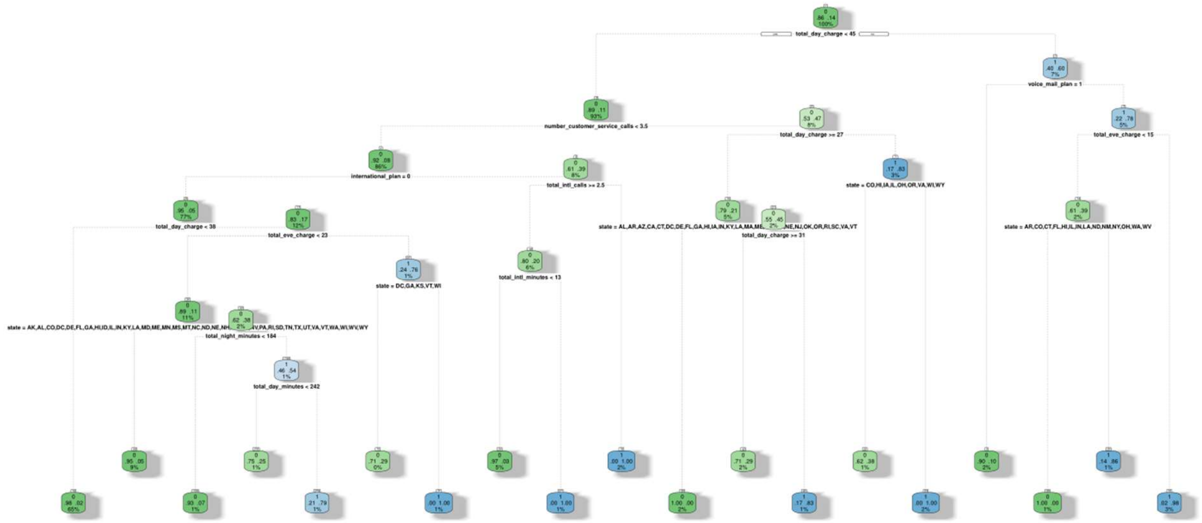
- When the dependent variable is binary, logistic regression is the suitable predictive analysis to undertake.

- Data are described and the link between a binary dependent variable and one or more nominal, ordinal, interval, or ratio-level independent variables are explained using logistic regression.
- Each variable is assumed to have a linear connection in logistic regression. The likelihood that a consumer will leave a service provider will, however, change dramatically if they are paying less for a given service than a rival.

## Decision Trees:



```
Classification tree:
rpart(formula = churn ~ ., data = churn_T_Train, method = "class")

Variables actually used in tree construction:
 [1] international_plan          number_customer_service_calls state
total_day_charge              total_day_minutes
 [6] total_eve_charge            total_intl_calls              total_intl_minutes
total_night_minutes           voice_mail_plan

Root node error: 266/1841 = 0.14449

n= 1841

        CP nsplit rel error  xerror     xstd
1 0.093985      0   1.00000 1.00000 0.056712
2 0.073308      2   0.81203 0.83835 0.052630
3 0.065789      4   0.66541 0.70677 0.048844
4 0.031955      7   0.45489 0.46992 0.040579
5 0.020677      9   0.39098 0.48872 0.041323
6 0.015038     11   0.34962 0.49248 0.041469
7 0.011278     14   0.30451 0.48496 0.041175
8 0.010025     15   0.29323 0.51504 0.042334
9 0.010000     18   0.26316 0.51504 0.042334
```


total_day_charge < 45
voice_mail_plan = 1
number_customer_service_calls < 3.5
total_day_charge >= 27
total_eve_charge < 15
international_plan = 0
total_intl_calls >= 2.5
state = CO,HI,IA,IL,OH,OR,VA,WI,WY
total_day_charge < 38
total_eve_charge < 23
state = AL,AR,AZ,CA,CT,DC,DE,FL,GA,HI,IA,IN,KY,LA,MA,ME,NE,NJ,OK,OR,RI,SC,VA,VT
total_day_charge >= 31
state = AR,CO,CT,FL,HI,IL,IN,LA,ND,NM,NY,OH,WA,WV
state = DC,GA,KS,VT,WI
total_intl_minutes < 13
state = AK,AL,CO,DC,DE,FL,GA,HI,ID,IL,IN,KY,LA,MD,ME,MN,MS,MT,NC,ND,NE,NH,NV,PA,RI,SD,TN,TX,UT,VA,VT,WA,WI,WV,WY
total_night_minutes < 184
total_day_minutes < 242

Reasons for using Decision Trees:

• The decision tree model is simple to comprehend, understand, and apply to both classification and regression issues. A decision tree's output is similarly simple to comprehend.

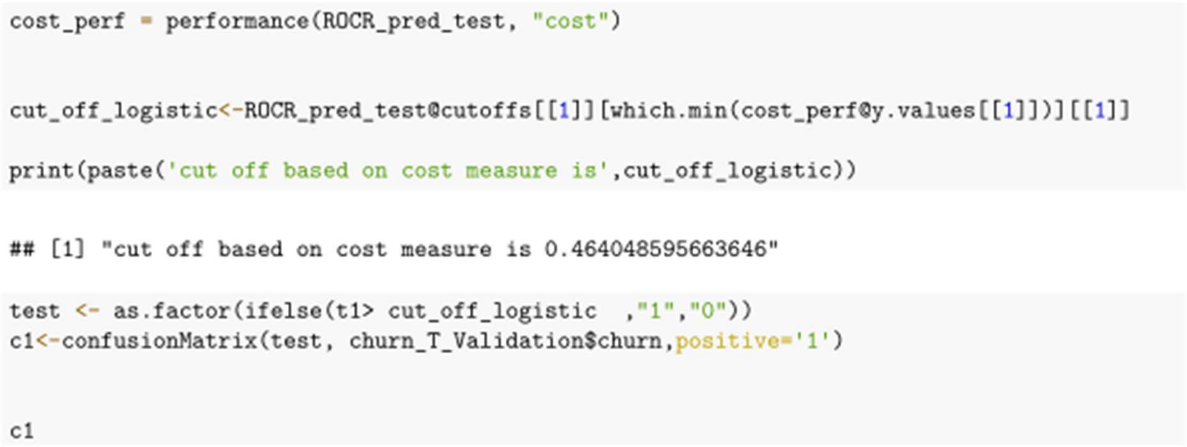
- A decision tree is one of the quickest ways to detect links between variables and the most significant variable.

- A decision tree needs less work during pre-processing than other algorithms and does not require normalization of data. For improved target variable prediction, new characteristics can also be developed.

## Estimation of model's performance:

### Logistic Regression:

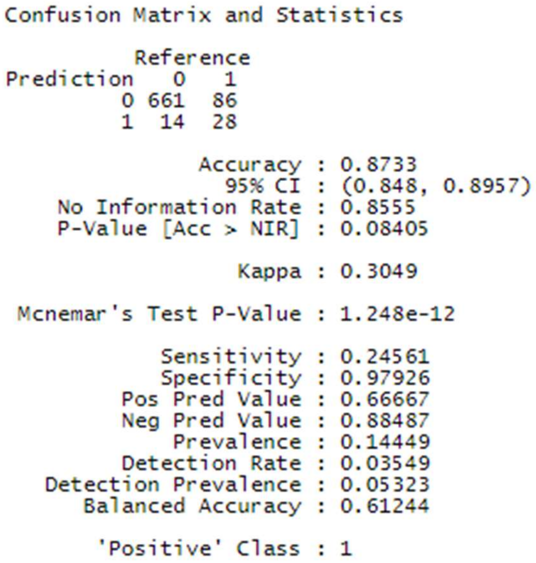
Built the model using train data set with logistic regression and predicted churn for the validation dataset. Based on the cost measure, we have determined the cutoff value to be 0.46404. Then compared those predicted values to the actual values and observed an accuracy of 87.33% with sensitivity of 24.56%. Also, we can observe that there are 100 miscalculations.



```
cost_perf = performance(ROCR_pred_test, "cost")


cut_off_logistic<-ROCR_pred_test@cutoffs[[1]][which.min(cost_perf@y.values[[1]])][[1]]

print(paste('cut off based on cost measure is',cut_off_logistic))
```

```
## [1] "cut off based on cost measure is 0.464048595663646"
```

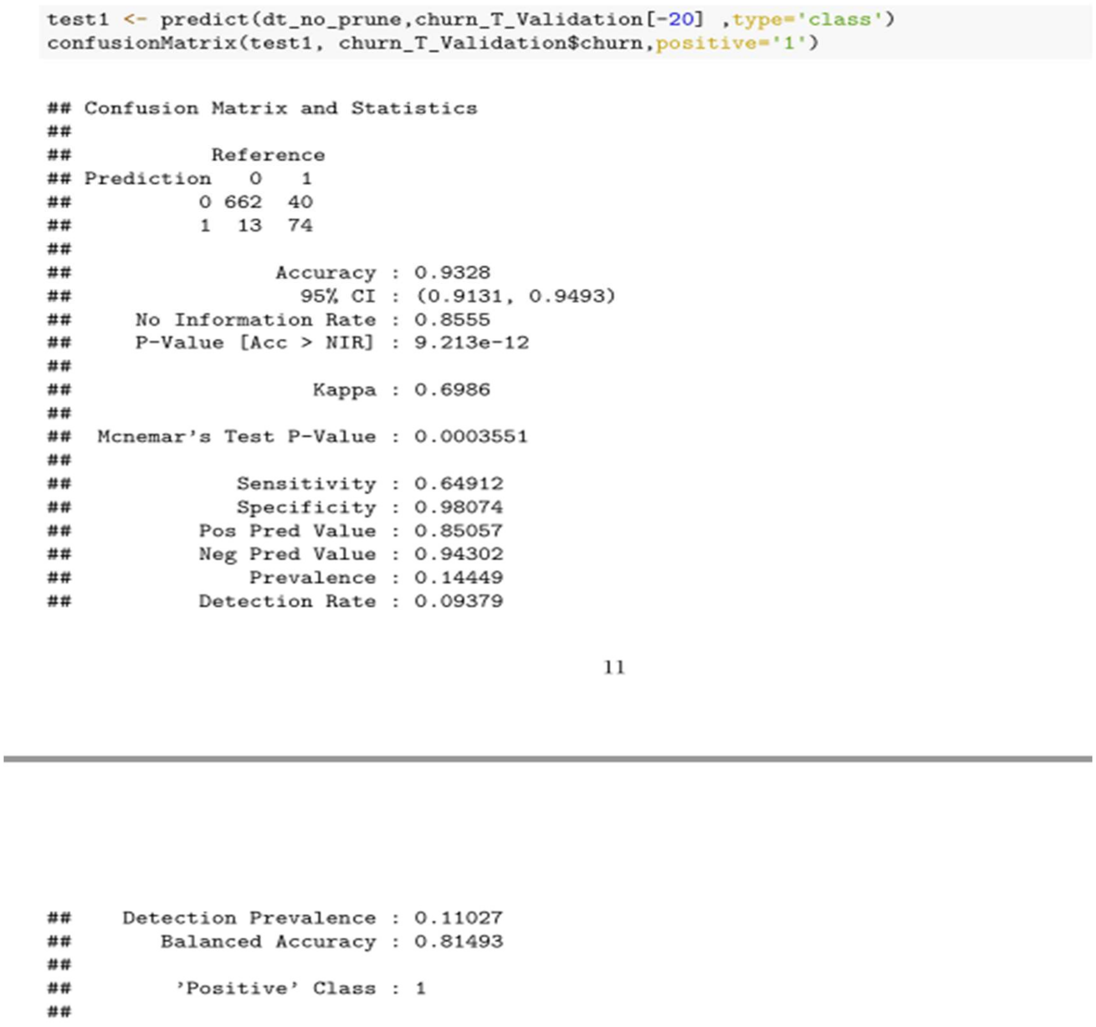
```
test <- as.factor(ifelse(t1> cut_off_logistic  ,"1","0"))
c1<-confusionMatrix(test, churn_T_Validation$churn,positive='1')


c1
```



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 661  86
         1  14  28

               Accuracy : 0.8733
                 95% CI : (0.848, 0.8957)
    No Information Rate : 0.8555
    P-Value [Acc > NIR] : 0.08405

                  Kappa : 0.3049

 Mcnemar's Test P-Value : 1.248e-12

            Sensitivity : 0.24561
            Specificity : 0.97926
         Pos Pred Value : 0.66667
         Neg Pred Value : 0.88487
             Prevalence : 0.14449
         Detection Rate : 0.03549
   Detection Prevalence : 0.05323
      Balanced Accuracy : 0.61244

       'Positive' Class : 1
```

## Decision Trees:

Implemented the same process as logistic regression to decision trees with default cutoff of 0.5. We found the accuracy of the model to be 93.28% and sensitivity of 64.9%.



```
test1 <- predict(dt_no_prune,churn_T_Validation[-20] ,type='class')
confusionMatrix(test1, churn_T_Validation$churn,positive='1')
```

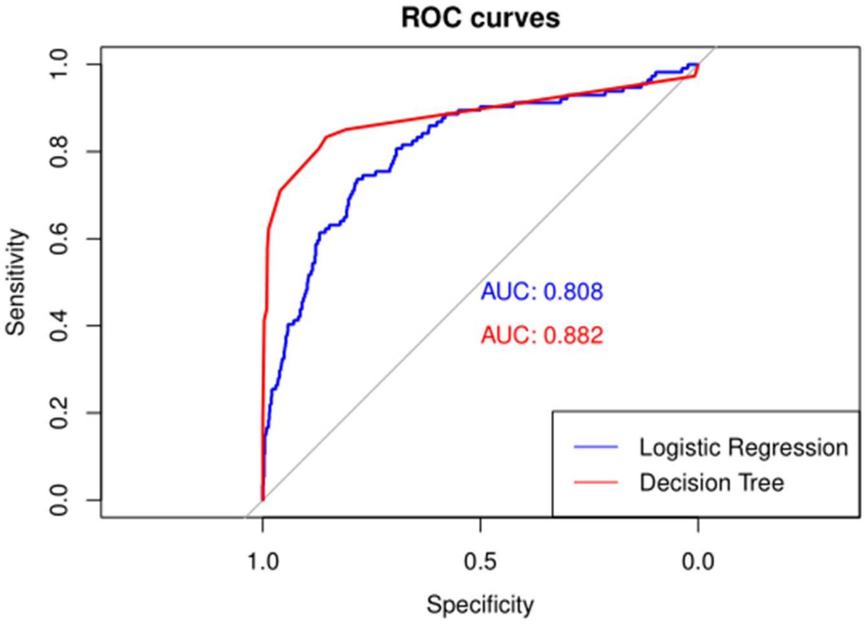
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 662  40
##          1  13  74
##
##                Accuracy : 0.9328
##                  95% CI : (0.9131, 0.9493)
##     No Information Rate : 0.8555
##     P-Value [Acc > NIR] : 9.213e-12
##
##                   Kappa : 0.6986
##
##  Mcnemar's Test P-Value : 0.0003551
##
##             Sensitivity : 0.64912
##             Specificity : 0.98074
##          Pos Pred Value : 0.85057
##          Neg Pred Value : 0.94302
##              Prevalence : 0.14449
##          Detection Rate : 0.09379
```

11

```
##    Detection Prevalence : 0.11027
##       Balanced Accuracy : 0.81493
##
##        'Positive' Class : 1
##
```

## Comparing Logistic Regression and Decision Trees:

- With logistic regression, we can see an accuracy of 87% with 24.5% sensitivity
- With decision tree model, we can see an accuracy of 93% with 64.9% sensitivity
- By Analyzing the performance of each model, below are the ROC curves of decision trees and logistic regression models.



- From the above graph, we can observe that Decision trees gives an AUC of 88% whereas Logistic regression model gives us an AUC of 80%
- Since the data is imbalanced and our priority is to choose the customers who are likely to churn. This can also be known as sensitivity. Hence, we are choosing decision tress because it has 64% of sensitivity with AUC of 88%

# Insights & Conclusions:

## Insights:

1. We can observe that the churn ratio in the data set is imbalanced. The data is spread over 86 to 14% of churn ratio which can be called as imbalanced dataset.
2. Based on the customer usage history, we can conclude that users tend to do international calls without an international plan, might be due to high rates in plans. This can be a significant variable in predicting churn.
3. There is a strong relation between user calls on day and evening. Users are either talking for more duration or for shorter duration.
4. It can also be observed that area 415 has large number of users with relatively high number of churn customers.
5. It is also observed that states Maryland and Texas have high count of churn customers.
6. These variables are significantly affecting the customers to churn.

## Conclusion:

We determined and developed recommendations that could help with the churn issue using the historical data and the study.

- Launch of price-competitive international calling packages.
- Lowering the per-minute rate would aid in keeping customers.
- Given how frequently the area code 415 is used, network coverage should be expanded in that area.
- Since most customer service calls range between 0 and 2, customer support should be more effective and deliver clients satisfactory solutions.
- Lowering price and implementing new plans may increase the cost for the company but there need to be trade off point where competitive pricing and providing seasonal offers to customers can make customers to stay in the company

-------------------------------------------------------------------------------------------

*Individual Student's Contribution to the project:*

Karthik Badiganti – Model Building Techniques, Variable selections, and comparison analysis between models. Insights and recommendations

**Tejaswini Yeruva** -Pruning the Decision Tree, performance measures and data transformation.

**Nikhil Reddy Addula, Venu Dodda–** Data Loading and Cleaning, Univariate Analysis of the data, Data interpretation with the variables and analysis.

-------------------------------------------------------------------------------------------