Business Analytics Final Project: Group-3

Rohith Chandra Koyyala, Manaswini Purumandla, Tejasvini Mavuleti , Riba Khan

2022-05-11

# Churn prediction

Customers’ telecom data from ABC Wireless Telecom is described in the Churn data dataset. There are 20 distinct characteristics. There are numerical and category variables in this dataset. The goal variable is churn, while the remaining 19 variables are predictors.

1. state (categorical)
2. account\_length
3. area\_code
4. international\_plan (yes/no)
5. voice\_mail\_plan (yes/no)
6. number\_vmail\_messages
7. total\_day\_minutes
8. total\_day\_calls
9. total\_day\_charge
10. total\_eve\_minutes
11. total\_eve\_calls
12. total\_eve\_charge
13. total\_night\_minutes
14. total\_night\_calls
15. total\_night\_charge
16. total\_intl\_minutes
17. total\_intl\_calls
18. total\_intl\_charge
19. number\_customer\_service\_calls.
20. Churn- The variable that you need to predict (target variable) is churn which takes two values ‘no’ and ‘yes’.

Installing librarys

library(ISLR)  
library(caret)  
library(class)  
library(pROC)  
library(rpart)  
library(dplyr)  
library(tidyr)  
library(ggplot2)  
library(ggcorrplot)  
library(rattle)  
library(rpart.plot)  
#install.packages("mice")  
library(mice)  
#install.packages("ranger")  
library(ranger)  
#install.packages("party")  
library("party")

Loading the data

churndata <- read.csv("C:/Users/kramr/Downloads/Churn\_Train.csv")

##Data exploration

summery of data set

summary(churndata)

## state account\_length area\_code international\_plan  
## Length:3333 Min. :-209.00 Length:3333 Length:3333   
## Class :character 1st Qu.: 72.00 Class :character Class :character   
## Mode :character Median : 100.00 Mode :character Mode :character   
## Mean : 97.32   
## 3rd Qu.: 127.00   
## Max. : 243.00   
## NA's :501   
## voice\_mail\_plan number\_vmail\_messages total\_day\_minutes total\_day\_calls  
## Length:3333 Min. :-10.000 Min. : 0.0 Min. : 0.0   
## Class :character 1st Qu.: 0.000 1st Qu.: 149.3 1st Qu.: 87.0   
## Mode :character Median : 0.000 Median : 190.5 Median :101.0   
## Mean : 7.333 Mean : 418.9 Mean :100.3   
## 3rd Qu.: 16.000 3rd Qu.: 237.8 3rd Qu.:114.0   
## Max. : 51.000 Max. :2185.1 Max. :165.0   
## NA's :200 NA's :200 NA's :200   
## total\_day\_charge total\_eve\_minutes total\_eve\_calls total\_eve\_charge  
## Min. : 0.00 Min. : 0.0 Min. : 0.0 Min. : 0.00   
## 1st Qu.:24.45 1st Qu.: 170.5 1st Qu.: 87.0 1st Qu.:14.14   
## Median :30.65 Median : 209.9 Median :100.0 Median :17.09   
## Mean :30.63 Mean : 324.3 Mean :100.1 Mean :17.08   
## 3rd Qu.:36.84 3rd Qu.: 257.6 3rd Qu.:114.0 3rd Qu.:20.00   
## Max. :59.64 Max. :1244.2 Max. :170.0 Max. :30.91   
## NA's :200 NA's :301 NA's :200 NA's :200   
## total\_night\_minutes total\_night\_calls total\_night\_charge total\_intl\_minutes  
## Min. : 23.2 Min. : 33.0 Min. : 1.040 Min. : 0.00   
## 1st Qu.:167.3 1st Qu.: 87.0 1st Qu.: 7.530 1st Qu.: 8.50   
## Median :201.4 Median :100.0 Median : 9.060 Median :10.30   
## Mean :201.2 Mean :100.1 Mean : 9.054 Mean :10.23   
## 3rd Qu.:235.3 3rd Qu.:113.0 3rd Qu.:10.590 3rd Qu.:12.10   
## Max. :395.0 Max. :175.0 Max. :17.770 Max. :20.00   
## NA's :200 NA's :200 NA's :200   
## total\_intl\_calls total\_intl\_charge number\_customer\_service\_calls  
## Min. : 0.00 Min. :0.000 Min. :0.000   
## 1st Qu.: 3.00 1st Qu.:2.300 1st Qu.:1.000   
## Median : 4.00 Median :2.780 Median :1.000   
## Mean : 4.47 Mean :2.762 Mean :1.561   
## 3rd Qu.: 6.00 3rd Qu.:3.270 3rd Qu.:2.000   
## Max. :20.00 Max. :5.400 Max. :9.000   
## NA's :301 NA's :200 NA's :200   
## churn   
## Length:3333   
## Class :character   
## Mode :character   
##   
##   
##   
##

Data preparation

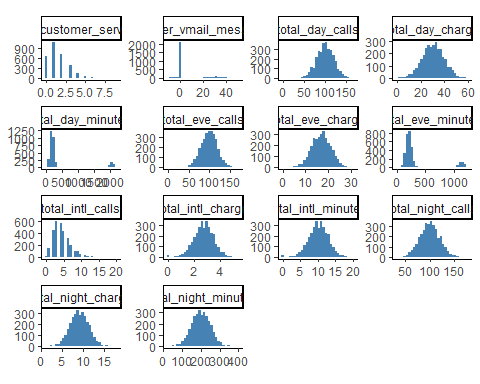
#Convert Categorical variables into factors  
churndata$state <- as.factor(churndata$state)  
churndata$area\_code <- as.factor(churndata$area\_code)  
churndata$international\_plan <- as.factor(churndata$international\_plan)  
churndata$voice\_mail\_plan <- as.factor(churndata$voice\_mail\_plan)  
churndata$churn <- as.factor(churndata$churn)  
  
churn\_true <- subset(churndata, churndata$churn == "yes")  
churn\_false <- subset(churndata, churndata$churn == "no")  
  
#No of churn count of yes/no   
churn\_count<-table(churndata$churn)

Data visualization

# Distribution of each variable in the dataset  
  
churndata[, 6:19] %>%   
 gather(key = Variable, value = Value) %>%   
 ggplot() +  
 geom\_histogram(aes(x = Value), fill = "steelblue") +  
 facet\_wrap(~Variable, scales='free') +  
 theme\_classic() +  
 theme(aspect.ratio = 0.5, axis.title = element\_blank(), panel.grid = element\_blank())

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 2802 rows containing non-finite values (stat\_bin).

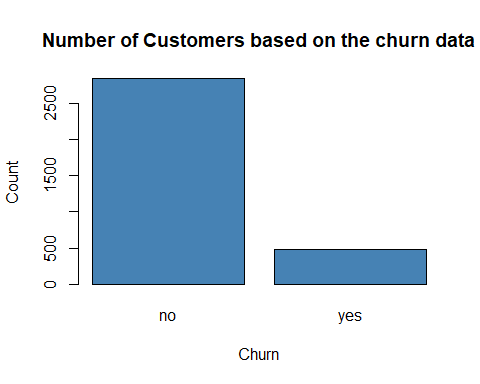


For the most part, we can see a flawless bell curve distribution of data.

There are a few anomalies in “Total day minutes” and “Total evening minutes.” Also, the statistics on “customer support calls” is biased.

# number of customers based on churn data

barplot(churn\_count,xlab ="Churn",ylab="Count" ,col = "steelblue" ,main = "Number of Customers based on the churn data")



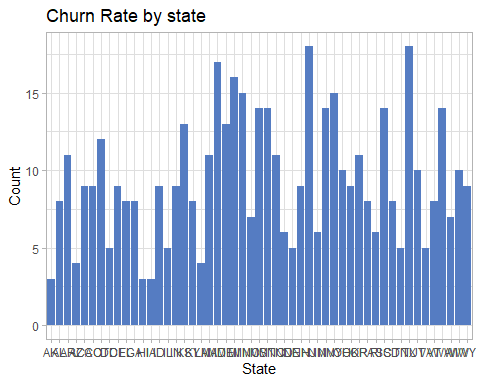
We can see that there were 2850 consumers who did not move and 483 customers who did switch.

# number of churn customers by state

Churn\_count\_state<-churn\_true %>% group\_by(state) %>% summarise(Churn\_state\_count=n())  
  
ChurnOnStates <- churndata %>% group\_by(churndata$state, churndata$churn) %>% summarise(count = n()) %>% mutate(prop = count / sum(count) \* 100)

## `summarise()` has grouped output by 'churndata$state'. You can override using  
## the `.groups` argument.

ggplot(Churn\_count\_state) +  
 aes(x = state, weight = Churn\_state\_count) +  
 geom\_bar(fill = "#557CC2") +  
 labs(x = "State", y = "Count", title = "Churn Rate by state") +  
 theme\_light()

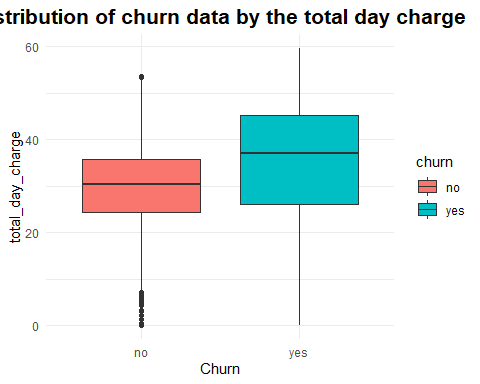


The states of Texas, New Jersey, Maryland, and Michigan have a high churn rate.

# Distribution of churn data by the total day charge

ggplot(churndata) +  
 aes(x = churn, y = total\_day\_charge, fill = churn) +  
 geom\_boxplot(shape = "circle") +  
 scale\_fill\_hue(direction = 1) +  
 labs(x = "Churn", y = "total\_day\_charge",title = "Distribution of churn data by the total day charge") +  
 theme\_minimal()+  
 theme(plot.title = element\_text(size = 16L,   
 face = "bold", hjust = 0.5))

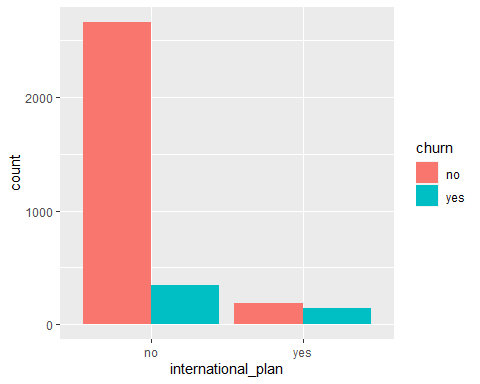
## Warning: Removed 200 rows containing non-finite values (stat\_boxplot).



Customers with day charges of 35 to 40 are more likely to cancel the service and switch to another provider, according to the above box plot distribution.

# Customers that had ‘international\_plan’ based on the churn data

ggplot(data = churndata, aes(x = international\_plan, y = ..count.., fill = churn)) +  
 geom\_bar(position = "dodge")



# 28% of all international plan subscribers switched.  
churn\_true %>%   
 group\_by(international\_plan) %>%   
 select(international\_plan) %>%   
 dplyr:: summarise("Churn Count" =n(), "Percent" = n()/483)

## # A tibble: 2 × 3  
## international\_plan `Churn Count` Percent  
## <fct> <int> <dbl>  
## 1 no 346 0.716  
## 2 yes 137 0.284

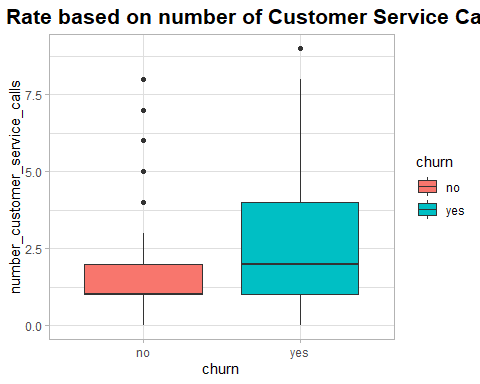
Few customers who had signed up for the overseas package had canceled it.

It is certain that 28% of all international plan users will cancel their subscriptions.

# Churn data based on Number of Customer service calls

ggplot(churndata) +  
 aes(x = churn, y = number\_customer\_service\_calls, fill = churn) +  
 geom\_boxplot(shape = "circle") +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = "Churn Rate based on number of Customer Service Calls") +  
 theme\_light() +  
 theme(plot.title = element\_text(size = 16L, face = "bold", hjust = 0.5))

## Warning: Removed 200 rows containing non-finite values (stat\_boxplot).



churn\_true %>%   
 filter(number\_customer\_service\_calls >= 1 & number\_customer\_service\_calls <= 4) %>%   
 tally()/483

## n  
## 1 0.6397516

# 64% of all customers who churned made 1 to 4 calls

Customers that phoned Customer Service more than 3 to 4 times are more likely to cancel the service, according to the above box plot distribution.

The graph above shows the number of service calls made by customers who had their services cancelled.

We can also notice that 64 percent of all churned consumers made 1 to 4 calls to the Customer Service center.

# Data Cleaning

# Handling NA Values - Imputing Missing values using mice package

set.seed(123)  
  
# As per Mice total\_night\_charge and total\_intl\_charge are multicolinear variables.So Mice will not impute the NAs for these columns. In order to make impute happen following steps are needed.  
churndata$total\_night\_charge[1] <- 2  
churndata$total\_intl\_charge[1] <- 0.5  
  
miceMod <- mice(churndata[, -20], method = "rf") #perform mice imputation, based on random forests.

##   
## iter imp variable  
## 1 1 account\_length number\_vmail\_messages total\_day\_minutes total\_day\_calls total\_day\_charge total\_eve\_minutes total\_eve\_calls total\_eve\_charge total\_night\_minutes total\_night\_charge total\_intl\_minutes total\_intl\_calls total\_intl\_charge number\_customer\_service\_calls  
## 1 2 account\_length number\_vmail\_messages total\_day\_minutes total\_day\_calls total\_day\_charge total\_eve\_minutes total\_eve\_calls total\_eve\_charge total\_night\_minutes total\_night\_charge total\_intl\_minutes total\_intl\_calls total\_intl\_charge number\_customer\_service\_calls  
## 1 3 account\_length number\_vmail\_messages total\_day\_minutes total\_day\_calls total\_day\_charge total\_eve\_minutes total\_eve\_calls total\_eve\_charge total\_night\_minutes total\_night\_charge total\_intl\_minutes total\_intl\_calls total\_intl\_charge number\_customer\_service\_calls  
## 1 4 account\_length number\_vmail\_messages total\_day\_minutes total\_day\_calls total\_day\_charge total\_eve\_minutes total\_eve\_calls total\_eve\_charge total\_night\_minutes total\_night\_charge total\_intl\_minutes total\_intl\_calls total\_intl\_charge number\_customer\_service\_calls  
## 1 5 account\_length number\_vmail\_messages total\_day\_minutes total\_day\_calls total\_day\_charge total\_eve\_minutes total\_eve\_calls total\_eve\_charge total\_night\_minutes total\_night\_charge total\_intl\_minutes total\_intl\_calls total\_intl\_charge number\_customer\_service\_calls  
## 2 1 account\_length number\_vmail\_messages total\_day\_minutes total\_day\_calls total\_day\_charge total\_eve\_minutes total\_eve\_calls total\_eve\_charge total\_night\_minutes total\_night\_charge total\_intl\_minutes total\_intl\_calls total\_intl\_charge number\_customer\_service\_calls  
## 2 2 account\_length number\_vmail\_messages total\_day\_minutes total\_day\_calls total\_day\_charge total\_eve\_minutes total\_eve\_calls total\_eve\_charge total\_night\_minutes total\_night\_charge total\_intl\_minutes total\_intl\_calls total\_intl\_charge number\_customer\_service\_calls  
## 2 3 account\_length number\_vmail\_messages total\_day\_minutes total\_day\_calls total\_day\_charge total\_eve\_minutes total\_eve\_calls total\_eve\_charge total\_night\_minutes total\_night\_charge total\_intl\_minutes total\_intl\_calls total\_intl\_charge number\_customer\_service\_calls  
## 2 4 account\_length number\_vmail\_messages total\_day\_minutes total\_day\_calls total\_day\_charge total\_eve\_minutes total\_eve\_calls total\_eve\_charge total\_night\_minutes total\_night\_charge total\_intl\_minutes total\_intl\_calls total\_intl\_charge number\_customer\_service\_calls  
## 2 5 account\_length number\_vmail\_messages total\_day\_minutes total\_day\_calls total\_day\_charge total\_eve\_minutes total\_eve\_calls total\_eve\_charge total\_night\_minutes total\_night\_charge total\_intl\_minutes total\_intl\_calls total\_intl\_charge number\_customer\_service\_calls  
## 3 1 account\_length number\_vmail\_messages total\_day\_minutes total\_day\_calls total\_day\_charge total\_eve\_minutes total\_eve\_calls total\_eve\_charge total\_night\_minutes total\_night\_charge total\_intl\_minutes total\_intl\_calls total\_intl\_charge number\_customer\_service\_calls  
## 3 2 account\_length number\_vmail\_messages total\_day\_minutes total\_day\_calls total\_day\_charge total\_eve\_minutes total\_eve\_calls total\_eve\_charge total\_night\_minutes total\_night\_charge total\_intl\_minutes total\_intl\_calls total\_intl\_charge number\_customer\_service\_calls  
## 3 3 account\_length number\_vmail\_messages total\_day\_minutes total\_day\_calls total\_day\_charge total\_eve\_minutes total\_eve\_calls total\_eve\_charge total\_night\_minutes total\_night\_charge total\_intl\_minutes total\_intl\_calls total\_intl\_charge number\_customer\_service\_calls  
## 3 4 account\_length number\_vmail\_messages total\_day\_minutes total\_day\_calls total\_day\_charge total\_eve\_minutes total\_eve\_calls total\_eve\_charge total\_night\_minutes total\_night\_charge total\_intl\_minutes total\_intl\_calls total\_intl\_charge number\_customer\_service\_calls  
## 3 5 account\_length number\_vmail\_messages total\_day\_minutes total\_day\_calls total\_day\_charge total\_eve\_minutes total\_eve\_calls total\_eve\_charge total\_night\_minutes total\_night\_charge total\_intl\_minutes total\_intl\_calls total\_intl\_charge number\_customer\_service\_calls  
## 4 1 account\_length number\_vmail\_messages total\_day\_minutes total\_day\_calls total\_day\_charge total\_eve\_minutes total\_eve\_calls total\_eve\_charge total\_night\_minutes total\_night\_charge total\_intl\_minutes total\_intl\_calls total\_intl\_charge number\_customer\_service\_calls  
## 4 2 account\_length number\_vmail\_messages total\_day\_minutes total\_day\_calls total\_day\_charge total\_eve\_minutes total\_eve\_calls total\_eve\_charge total\_night\_minutes total\_night\_charge total\_intl\_minutes total\_intl\_calls total\_intl\_charge number\_customer\_service\_calls  
## 4 3 account\_length number\_vmail\_messages total\_day\_minutes total\_day\_calls total\_day\_charge total\_eve\_minutes total\_eve\_calls total\_eve\_charge total\_night\_minutes total\_night\_charge total\_intl\_minutes total\_intl\_calls total\_intl\_charge number\_customer\_service\_calls  
## 4 4 account\_length number\_vmail\_messages total\_day\_minutes total\_day\_calls total\_day\_charge total\_eve\_minutes total\_eve\_calls total\_eve\_charge total\_night\_minutes total\_night\_charge total\_intl\_minutes total\_intl\_calls total\_intl\_charge number\_customer\_service\_calls  
## 4 5 account\_length number\_vmail\_messages total\_day\_minutes total\_day\_calls total\_day\_charge total\_eve\_minutes total\_eve\_calls total\_eve\_charge total\_night\_minutes total\_night\_charge total\_intl\_minutes total\_intl\_calls total\_intl\_charge number\_customer\_service\_calls  
## 5 1 account\_length number\_vmail\_messages total\_day\_minutes total\_day\_calls total\_day\_charge total\_eve\_minutes total\_eve\_calls total\_eve\_charge total\_night\_minutes total\_night\_charge total\_intl\_minutes total\_intl\_calls total\_intl\_charge number\_customer\_service\_calls  
## 5 2 account\_length number\_vmail\_messages total\_day\_minutes total\_day\_calls total\_day\_charge total\_eve\_minutes total\_eve\_calls total\_eve\_charge total\_night\_minutes total\_night\_charge total\_intl\_minutes total\_intl\_calls total\_intl\_charge number\_customer\_service\_calls  
## 5 3 account\_length number\_vmail\_messages total\_day\_minutes total\_day\_calls total\_day\_charge total\_eve\_minutes total\_eve\_calls total\_eve\_charge total\_night\_minutes total\_night\_charge total\_intl\_minutes total\_intl\_calls total\_intl\_charge number\_customer\_service\_calls  
## 5 4 account\_length number\_vmail\_messages total\_day\_minutes total\_day\_calls total\_day\_charge total\_eve\_minutes total\_eve\_calls total\_eve\_charge total\_night\_minutes total\_night\_charge total\_intl\_minutes total\_intl\_calls total\_intl\_charge number\_customer\_service\_calls  
## 5 5 account\_length number\_vmail\_messages total\_day\_minutes total\_day\_calls total\_day\_charge total\_eve\_minutes total\_eve\_calls total\_eve\_charge total\_night\_minutes total\_night\_charge total\_intl\_minutes total\_intl\_calls total\_intl\_charge number\_customer\_service\_calls

## Warning: Number of logged events: 350

miceOutput <- complete(miceMod) #generate the complete data  
anyNA(miceOutput)

## [1] FALSE

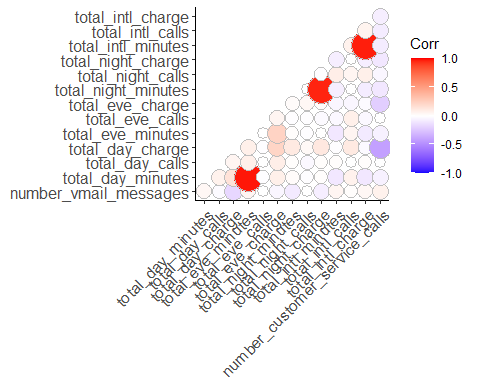
churndata\_Imputed <- mutate(miceOutput,churn=churndata$churn)  
#summary(churndata\_Imputed)

# Correlation Plots

When churn equals Yes, we’ll look at the correlation of variables:

#str(churndata)  
churn\_yes<-churndata\_Imputed %>% filter(churn=='yes')  
Corr\_churn\_cust<- cor(churn\_yes[, 6:19])   
# ggplot to determine the correlation between variables in the when churn = yes  
ggcorrplot(Corr\_churn\_cust, method = "circle", type = "lower", ggtheme = theme\_classic)

## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =  
## "none")` instead.



We discovered a high negative association between the number of customer support calls and total day, total evening, and total international costs among those who churned from the plots above.

We can deduce several potentially interesting truths from the correlation. When churn = yes, the higher the costs, the more calls to customer support were made.

# Model Selection:

To find the most accurate model for predicting which customers would churn and which will not.

A predictive model based on regression and Decision Tree Model was used to highlight the effect of numerous factors and their relevance in forecasting the result of the target variable.

Regression can be done in two ways:

* Linear Regression
* Logistic Regression

Because the data’s target variable is categorical, a logistic regression model is the best option. When predicting a binomial property, it’s tempting to use linear regression as a model, however the performance likelihood might be negative or more than 1, making it ineffective. A likelihood or chance of odds between 0 and 1, as determined by logistic regression, is the optimum outcome for this model.

We also picked Logistic Regression and Decision Models as appropriate after examining the dataset because categorization is our primary goal. In that vein, we’ll test both models on our dataset and choose the best one to predict the test dataset as the final model.

# Determining the predictive ability of Logistic regression and Decision trees models :

To prevent overfitting the model, divide the dataset into two sections: training and validation. Building a logistic regression model, predicting the outcomes on the validation data set, and validating the model’s performance with a confusion matrix Building a Decision Tree Model and Predicting the Results on the Validation Data Set and Validating the Model’s Performance with the Confusion Matrix Compare the Model Performance and Chose the best model.

# Data Partitioning

set.seed(123)  
index<- createDataPartition(churndata\_Imputed$churn,p=0.8,list=FALSE)  
  
train\_data<-churndata\_Imputed[index,]  
validation\_data <- churndata\_Imputed[-index,]

# Building Logistic Regression Model

**Logistic Regression**

Logistic Regression - Logistic regression is a type of regression that uses a combination of continuous and discrete factors to predict discrete or categorical variables. To put it another way, the Y or goal variables must always be categorical variables, but the X variables might be categorical or continuous variables.

set.seed(123)  
Logistic\_Model <- glm(churn~.,data=train\_data ,family = "binomial" )  
#summary(Logistic\_Model)  
  
predict\_validation<-predict(Logistic\_Model,validation\_data,type="response")  
head(predict\_validation)

## 3 5 6 9 12 21   
## 0.10932402 0.25050315 0.05955310 0.01219066 0.50233997 0.07534092

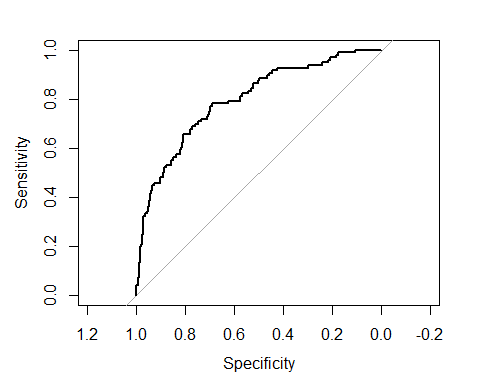
resultcheck1<-ifelse(predict\_validation>0.5,'yes','no')  
  
#Accuracy Check  
error1<-mean(resultcheck1!=validation\_data$churn)  
accuracy1<-1-error1  
print(accuracy1)

## [1] 0.8678679

plot.roc(validation\_data$churn,predict\_validation)

## Setting levels: control = no, case = yes

## Setting direction: controls < cases



# Confusion Matrix of Logistic Regression Model

set.seed(123)  
  
Logistic\_CM <- confusionMatrix(as.factor(resultcheck1),as.factor(validation\_data$churn))  
Logistic\_CM

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction no yes  
## no 556 74  
## yes 14 22  
##   
## Accuracy : 0.8679   
## 95% CI : (0.8398, 0.8927)  
## No Information Rate : 0.8559   
## P-Value [Acc > NIR] : 0.2052   
##   
## Kappa : 0.2764   
##   
## Mcnemar's Test P-Value : 3.187e-10   
##   
## Sensitivity : 0.9754   
## Specificity : 0.2292   
## Pos Pred Value : 0.8825   
## Neg Pred Value : 0.6111   
## Prevalence : 0.8559   
## Detection Rate : 0.8348   
## Detection Prevalence : 0.9459   
## Balanced Accuracy : 0.6023   
##   
## 'Positive' Class : no   
##

The Accuracy of the logistic Regression Model is 86.49 %

Sensitivity : 97.54%

Specificity : 20.83%

# Building Decision Tree model

#Decision Tree Model:

The main goal of the decision tree model is to Classify or predict an outcome based on a set of predictors.

set.seed(123)  
# Decision Tree  
  
DT\_model<- rpart(churn ~ .,data=train\_data,method = 'class')  
  
# Show the variable importance  
  
#DT\_model$variable.importance  
  
# Show the split for variable  
  
head(DT\_model$splits)

## count ncat improve index adj  
## total\_day\_charge 2667 -1 70.74761 45.125 0  
## number\_customer\_service\_calls 2667 -1 59.20504 3.500 0  
## international\_plan 2667 2 39.37377 1.000 0  
## total\_day\_minutes 2667 -1 21.03222 225.050 0  
## state 2667 51 12.99175 2.000 0  
## number\_customer\_service\_calls 2499 -1 60.13966 3.500 0

#Predict the probability  
Prob\_DT <- predict(DT\_model, newdata = validation\_data, type = "prob")  
  
  
#AUC Value  
roc(validation\_data$churn,Prob\_DT[,2])

## Setting levels: control = no, case = yes

## Setting direction: controls < cases

##   
## Call:  
## roc.default(response = validation\_data$churn, predictor = Prob\_DT[, 2])  
##   
## Data: Prob\_DT[, 2] in 570 controls (validation\_data$churn no) < 96 cases (validation\_data$churn yes).  
## Area under the curve: 0.8297

# Confusion Matrix of Decision Tree

set.seed(123)  
class\_decision\_tree <- predict(DT\_model, newdata = validation\_data, type = "class")  
  
confusionMatrix(as.factor(class\_decision\_tree),as.factor(validation\_data$churn))

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction no yes  
## no 552 37  
## yes 18 59  
##   
## Accuracy : 0.9174   
## 95% CI : (0.8939, 0.9372)  
## No Information Rate : 0.8559   
## P-Value [Acc > NIR] : 8.774e-07   
##   
## Kappa : 0.6353   
##   
## Mcnemar's Test P-Value : 0.01522   
##   
## Sensitivity : 0.9684   
## Specificity : 0.6146   
## Pos Pred Value : 0.9372   
## Neg Pred Value : 0.7662   
## Prevalence : 0.8559   
## Detection Rate : 0.8288   
## Detection Prevalence : 0.8844   
## Balanced Accuracy : 0.7915   
##   
## 'Positive' Class : no   
##

The Accuracy of the Decision Tree Model is 92.04 %

Sensitivity : 97.02%

Specificity : 62.50%

# Model Performance:

Because of its excellent accuracy, we picked the Decision Tree Model to forecast the churn of the test data after testing the Model’s performance.

# Building the Final Model to predict the churn using Test data and Decision Tree Algorithm

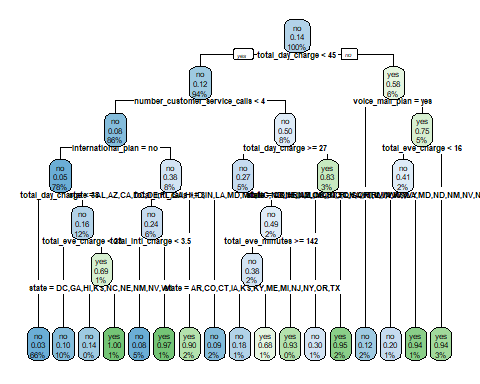
set.seed(123)  
  
#After testing for accuracy using validation and training data ,we can use the total dataset for building the actual model to predict the churn  
Model\_ABC\_Wireless<- rpart(churn ~ .,data=churndata\_Imputed,method = 'class')  
  
# Show the variable importance  
  
Model\_ABC\_Wireless$variable.importance

## total\_day\_charge state   
## 139.920014 84.068374   
## number\_customer\_service\_calls total\_eve\_charge   
## 82.550272 53.891680   
## international\_plan total\_intl\_charge   
## 53.720359 51.467303   
## total\_intl\_minutes total\_intl\_calls   
## 49.739629 42.239384   
## total\_day\_minutes number\_vmail\_messages   
## 41.654025 40.725792   
## voice\_mail\_plan total\_eve\_minutes   
## 38.993936 32.954992   
## total\_night\_calls total\_night\_charge   
## 5.823237 5.186486   
## total\_night\_minutes account\_length   
## 4.244571 3.018568   
## area\_code total\_day\_calls   
## 2.889524 2.591913

# Show the split for variable  
  
head(Model\_ABC\_Wireless$splits)

## count ncat improve index adj  
## total\_day\_charge 3333 -1 86.35605 44.975 0  
## number\_customer\_service\_calls 3333 -1 77.21135 3.500 0  
## international\_plan 3333 2 55.77483 1.000 0  
## total\_day\_minutes 3333 -1 24.95119 223.250 0  
## state 3333 51 14.95004 2.000 0  
## number\_customer\_service\_calls 3119 -1 80.74432 3.500 0

#Plot of DT  
  
#fancyRpartPlot(Model\_ABC\_Wireless)  
  
rpart.plot(Model\_ABC\_Wireless, cex=0.5)



#Predict the probability  
Prob\_decision\_tree <- predict(Model\_ABC\_Wireless, newdata = churndata\_Imputed, type = "prob")  
  
  
#AUC Value  
roc(churndata\_Imputed$churn,Prob\_decision\_tree[,2])

## Setting levels: control = no, case = yes

## Setting direction: controls < cases

##   
## Call:  
## roc.default(response = churndata\_Imputed$churn, predictor = Prob\_decision\_tree[, 2])  
##   
## Data: Prob\_decision\_tree[, 2] in 2850 controls (churndata\_Imputed$churn no) < 483 cases (churndata\_Imputed$churn yes).  
## Area under the curve: 0.8931

# Prediction of the Test data

set.seed(123)  
load("C:/Users/kramr/Downloads/Customers\_To\_Predict.RData")  
  
count(Customers\_To\_Predict)

## # A tibble: 1 × 1  
## n  
## <int>  
## 1 1600

#summary(Customers\_To\_Predict)  
  
# Check for NA Values  
#colMeans(is.na(Customers\_To\_Predict))  
  
Churn\_Prob <- predict(Model\_ABC\_Wireless,Customers\_To\_Predict,type = "prob")  
  
head(Churn\_Prob)

## no yes  
## 1 0.96969697 0.03030303  
## 2 0.96969697 0.03030303  
## 3 0.96969697 0.03030303  
## 4 0.92485549 0.07514451  
## 5 0.96969697 0.03030303  
## 6 0.04819277 0.95180723

predict\_churn <- predict(Model\_ABC\_Wireless,Customers\_To\_Predict,type = "class")  
head(predict\_churn)

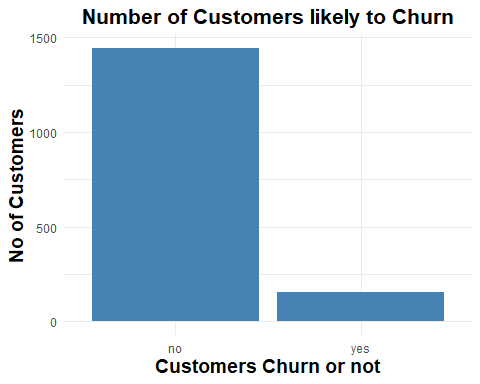
## 1 2 3 4 5 6   
## no no no no no yes   
## Levels: no yes

predict\_churn<- as.data.frame(predict\_churn)  
  
summary(predict\_churn)

## predict\_churn  
## no :1445   
## yes: 155

# Plot for summary of the Test data

ggplot(predict\_churn) +  
 aes(x = predict\_churn) +  
 geom\_bar(fill = "steelblue") +  
 labs(x = "Customers Churn or not",   
 y = "No of Customers", title = "Number of Customers likely to Churn") +  
 theme\_minimal() +  
 theme(plot.title = element\_text(size = 16L,   
 face = "bold", hjust = 0.5), axis.title.y = element\_text(size = 14L, face = "bold"), axis.title.x = element\_text(size = 14L,   
 face = "bold"))



predict\_churn

no :1444 yes: 156

# Conclusion:

From the data exploration,

Consumers who phoned Customer Account more than 2 to 4 times are more likely to cancel the service, implying that customers who moved companies were dissatisfied with the service.

Customers who paid higher (approx. above 35) day rates are more likely to discontinue the service, we may deduce.

International day costs, like day charges, have an impact on the turnover rate. Customers that spent greater (roughly. above 30)international day costs are more likely to terminate the service, according to the above box plot distribution.

# Recommendations to Curtail the Churn Rate:

Reduce the Day Time and International Day Time Charges or maintain the competitive charges on these two categories.

Improve and provide excellent the customer service

Overall quality should be maintained or improved (Bandwidth in highly dense areas etc)

Customers who are loyal should be rewarded in order to keep them.

Customer input should be sought on a regular basis.