TELCO CUSTOMER CHURN PREDICTION

ADEPOJU BOLATITO

6/18/2021

TABLE OF CONTENT

OVERVIEW

Problem Statement

Objective

METHODOLOGY

Data Preparation  
Data Inspection  
Data Cleaning  
Exploratory Data Analysis

Statistical Analysis

Model Building

Model Selection

Conclusion OVERVIEW

Problem Statement:The liberalization of the telecommunication market in Europe has led to significant customer churn, Consequently, Telco telecommunication company has partnered with DataLab Nigeria to help diagnose the cause of churning among customers.One of the hypotheses under consideration is that churn is driven by the customers’ price sensitivities,hence, the need to explore the data in order to validate or negate the hypothesis and also help to build a model that would help predict the likelihood of churning in future customers.

Aim: To understand the reason customer churned and also predict the likelihood of churning in future customers based on the data available using various machine learning models and deploying the best model.

Tools: R,tidyverse,ggbplot,Machine Learning

Models (LogisticRegression, randomforest).

METHODOLOGY

STEP 1.0

Data Preparation; This stage involves importing of the necessary data science libraries and the given dataset into R script using the imported libraries as shown below.

## Loading required package: tidyverse

## Warning: package 'tidyverse' was built under R version 4.0.5

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.3 v purrr 0.3.4  
## v tibble 3.1.1 v dplyr 1.0.6  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.1

## Warning: package 'ggplot2' was built under R version 4.0.5

## Warning: package 'tibble' was built under R version 4.0.5

## Warning: package 'tidyr' was built under R version 4.0.5

## Warning: package 'readr' was built under R version 4.0.5

## Warning: package 'purrr' was built under R version 4.0.5

## Warning: package 'dplyr' was built under R version 4.0.5

## Warning: package 'stringr' was built under R version 4.0.5

## Warning: package 'forcats' was built under R version 4.0.5

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

## Loading in and inspecting the dataset

##   
## -- Column specification --------------------------------------------------------  
## cols(  
## .default = col\_character(),  
## SeniorCitizen = col\_double(),  
## tenure = col\_double(),  
## MonthlyCharges = col\_double(),  
## TotalCharges = col\_double()  
## )  
## i Use `spec()` for the full column specifications.

## Warning: 1 parsing failure.  
## row col expected actual file  
## 4245 tenure a double #NAME? 'C:\Users\USER\Downloads\Telcom Dataset.csv'

## Rows: 7,043  
## Columns: 21  
## $ customerID <chr> "7590-VHVEG", "5575-GNVDE", "3668-QPYBK", "7795-CFOCW~  
## $ gender <chr> "Female", "Male", "Male", "Male", "Female", "Female",~  
## $ SeniorCitizen <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~  
## $ Partner <chr> "Yes", "No", "No", "No", "No", "No", "No", "No", "Yes~  
## $ Dependents <chr> "No", "No", "No", "No", "No", "No", "Yes", "No", "No"~  
## $ tenure <dbl> 1, 34, 2, 45, 2, 8, 22, 10, 28, 62, 13, 16, 58, 49, 2~  
## $ PhoneService <chr> "No", "Yes", "Yes", "No", "Yes", "Yes", "Yes", "No", ~  
## $ MultipleLines <chr> "No phone service", "No", "No", "No phone service", "~  
## $ InternetService <chr> "DSL", "DSL", "DSL", "DSL", "Fiber optic", "Fiber opt~  
## $ OnlineSecurity <chr> "No", "Yes", "Yes", "Yes", "No", "No", "No", "Yes", "~  
## $ OnlineBackup <chr> "Yes", "No", "Yes", "No", "No", "No", "Yes", "No", "N~  
## $ DeviceProtection <chr> "No", "Yes", "No", "Yes", "No", "Yes", "No", NA, "Yes~  
## $ TechSupport <chr> "No", "No", "No", "Yes", "No", "No", "No", "No", "Yes~  
## $ StreamingTV <chr> "No", "No", "No", "No", "No", "Yes", NA, "No", "Yes",~  
## $ StreamingMovies <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "Yes~  
## $ Contract <chr> "Month-to-month", "One year", "Month-to-month", "One ~  
## $ PaperlessBilling <chr> "Yes", "No", "Yes", "No", "Yes", "Yes", "Yes", "No", ~  
## $ PaymentMethod <chr> "Electronic check", NA, "Mailed check", "Bank transfe~  
## $ MonthlyCharges <dbl> 29.85, 56.95, 53.85, 42.30, 70.70, 99.65, 89.10, 29.7~  
## $ TotalCharges <dbl> 29.85, 1889.50, 108.15, 1840.75, 151.65, 820.50, 1949~  
## $ Churn <chr> "No", "No", "Yes", "No", "Yes", "Yes", "No", "No", "Y~

Tidyverse is the inbuilt library in R that is used for data inspection and data wrangling. the Telecom dataset has 21 columns and 7043 observations. it has 3 numerical columns while the remaining columns are character data type.

STEP 2.0

Data Inspection; This is the stage where the dataset is being checked for normality and abnormality. The number of features and observations are checked in this stage, the data types, presence of missing values and so on.

## # A tibble: 20 x 21  
## customerID gender SeniorCitizen Partner Dependents tenure PhoneService  
## <chr> <chr> <dbl> <chr> <chr> <dbl> <chr>   
## 1 7590-VHVEG Female 0 Yes No 1 No   
## 2 5575-GNVDE Male 0 No No 34 Yes   
## 3 3668-QPYBK Male 0 No No 2 Yes   
## 4 7795-CFOCW Male 0 No No 45 No   
## 5 9237-HQITU Female 0 No No 2 Yes   
## 6 9305-CDSKC Female 0 No No 8 Yes   
## 7 1452-KIOVK Male 0 No Yes 22 Yes   
## 8 6713-OKOMC Female 0 No No 10 No   
## 9 7892-POOKP Female 0 Yes No 28 Yes   
## 10 6388-TABGU Male 0 No Yes 62 Yes   
## 11 9763-GRSKD Male 0 Yes Yes 13 Yes   
## 12 7469-LKBCI Male 0 No No 16 Yes   
## 13 8091-TTVAX Male 0 Yes No 58 Yes   
## 14 0280-XJGEX Male 0 No No 49 Yes   
## 15 5129-JLPIS Male 0 No No 25 Yes   
## 16 3655-SNQYZ Female 0 Yes Yes 1200 Yes   
## 17 8191-XWSZG Female 0 No No 52 Yes   
## 18 9959-WOFKT Male 0 No Yes 71 Yes   
## 19 4190-MFLUW Female 0 Yes Yes 10 Yes   
## 20 4183-MYFRB Female 0 No No 21 Yes   
## # ... with 14 more variables: MultipleLines <chr>, InternetService <chr>,  
## # OnlineSecurity <chr>, OnlineBackup <chr>, DeviceProtection <chr>,  
## # TechSupport <chr>, StreamingTV <chr>, StreamingMovies <chr>,  
## # Contract <chr>, PaperlessBilling <chr>, PaymentMethod <chr>,  
## # MonthlyCharges <dbl>, TotalCharges <dbl>, Churn <chr>

## [1] 56

#### checking the unique entries in each column

## [1] "Female" "Male"

The Customers were either Male or female

## [1] 0 1

the customers were both senior citizens and non Senior citizen

## [1] "Yes" "No" NA

## [1] 1

some of the customers had partners while some were single. Although there was a response whose status wasnt known.it would be dealt with as a missing value later.

## [1] "No" "Yes" NA

## [1] 1

some of the customers had dependents while some didnt. there were also some customers with no answer.Only one customer had an unknown responsibility status.

## [1] 1 34 2 45 8 22 10 28 62 13 16 58 49 25 1200  
## [16] 52 71 21 12 30 47 72 17 27 5 46 11 70 63 69  
## [31] 43 15 60 18 66 9 3 31 50 64 56 7 42 35 48  
## [46] 29 65 38 68 32 55 37 36 41 6 4 33 67 23 57  
## [61] 61 14 20 53 40 59 24 44 19 54 51 26 0 39 NA  
## [76] 205

## [1] 4

The figures above shows how long the customers had stayed with the company. ranging from 1 month to 205 months. while there are also 4 customers with unknown tenure.

## [1] "No" "Yes"

Not all of the customers subscribed to Phone service. some used the phone service while others didnt.

## [1] "No phone service" "No" NA "Yes"

## [1] "NO PHONE SERVICE" "NO" NA "YES"

## [1] 4

Some of the customers had multiple lines while some had just a single line.four customers didn’t specify how many lines they had.

## [1] "DSL" "Fiber optic" "No" NA "dsl"

stringr is the library that help to make corrections to characters

## [1] "DSL" "FIBER OPTIC" "NO" NA

four customers didnt specify if they used internet service or not

## [1] 4

some of the customers didnt subscribe to internet service while others subscribed to different internet service types(DSL and fiber optic). we cant tell from the data if about 4 customers subscribed to internet service or not.

## [1] "No" "Yes" "No internet service"  
## [4] "NO"

## [1] "NO" "YES" "NO INTERNET SERVICE"

## [1] "NO" "YES"

some of the customers didn’t subscribe for internet service,hence no online security service. meanwhile some people enjoyed online security

## [1] "Yes" "No" "No internet service"  
## [4] NA "NO"

## [1] "YES" "NO" "NO INTERNET SERVICE"  
## [4] NA

## [1] "YES" "NO" NA

## [1] 3

same as the online security, customers that didn’t subscribe to internet service couldn’t have access to online backup,some had while we couldn’t get the status of 3 customers.

## [1] "No" "Yes" NA   
## [4] "No internet service"

## [1] "NO" "YES" NA

## [1] 5

some of the customers never subscribed to the device protection service while others did.however, about the status of 5 customers wasn’t specified.

## [1] "No" "Yes" "No internet service"  
## [4] NA "YES" "\\"

## [1] "NO" "YES" NA "\\"

## [1] 3

## [1] "NO" "YES" NA

## [1] 4

some of the customers had Tech support while others didn’t. meanwhile, the data didn’t provide if 4 customers had tech support or not.

## [1] "No" "Yes" NA   
## [4] "No internet service"

## [1] "NO" "YES" NA

## [1] 4

From the data,we can see that some of the customers streamed TV while some didn’t. Additionally, the data didnt indicate if 4 of the customers streamed TV or not.

## [1] "No" "Yes" NA   
## [4] "No internet service" "NO"

## [1] "NO" "YES" NA

## [1] 2

while some of the customers used their data to stream movies online, some didn’t.however some people couldn’t stream movies because they didn’t subscribe to internet service. although, the data didn’t provide if 2 of the customers streamed movies or not.

## [1] "Month-to-month" "One year" "Two year" NA   
## [5] "2 years"

## [1] "Month-to-month" "One year" "Two year" NA

## [1] 4

The customers activated different contract type with Telco communications, while some were paying month to month, some were in a yearly contract while others were in a two-year contract.however, the contract type of 4 of the customers wasn’t provided in the data.

## [1] "Yes" "No" NA

## [1] 2

some of the customers were billed on paper while some did paperless billing. the type of billing of 2 of the customers wasn’t known.

## [1] "Electronic check" NA   
## [3] "Mailed check" "Bank transfer (automatic)"  
## [5] "Credit card (automatic)" "Electronic\_check"

## [1] "Electronic check" NA   
## [3] "Mailed check" "Bank transfer (automatic)"  
## [5] "Credit card (automatic)"

## [1] 5

While some people were paying with Electronic check, some used credit card, Mailed check,bank transfer and credit card.even though the mode of payment pd 5 of the customers wasn’t specified.

## [1] 2

the figures above represents the monthly charge of the customers.we can see from the data that some of the customers were charged higher than the other. some customers were charged as low as 18.95 monthly while some went as high as 115. meanwhile, some people’s monthly charge wasn’t provided in the data

## [1] 12

The total charges of the customer as at the tome the data was collected is represented above. while the total charges of some customers wasn’t provided, some of the customers paid as low as 20.40 while some paid more than 8000.00

## [1] "No" "Yes"

## customerID gender SeniorCitizen Partner   
## Length:7043 Length:7043 Min. :0.0000 Length:7043   
## Class :character Class :character 1st Qu.:0.0000 Class :character   
## Mode :character Mode :character Median :0.0000 Mode :character   
## Mean :0.1621   
## 3rd Qu.:0.0000   
## Max. :1.0000   
##   
## Dependents tenure PhoneService MultipleLines   
## Length:7043 Min. : 0.00 Length:7043 Length:7043   
## Class :character 1st Qu.: 9.00 Class :character Class :character   
## Mode :character Median : 29.00 Mode :character Mode :character   
## Mean : 32.57   
## 3rd Qu.: 55.00   
## Max. :1200.00   
## NA's :4   
## InternetService OnlineSecurity OnlineBackup DeviceProtection   
## Length:7043 Length:7043 Length:7043 Length:7043   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## TechSupport StreamingTV StreamingMovies Contract   
## Length:7043 Length:7043 Length:7043 Length:7043   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## PaperlessBilling PaymentMethod MonthlyCharges TotalCharges   
## Length:7043 Length:7043 Min. : 18.25 Min. : 18.8   
## Class :character Class :character 1st Qu.: 35.50 1st Qu.: 401.7   
## Mode :character Mode :character Median : 70.35 Median :1397.7   
## Mean : 64.76 Mean :2283.6   
## 3rd Qu.: 89.85 3rd Qu.:3795.0   
## Max. :118.75 Max. :8684.8   
## NA's :2 NA's :12   
## Churn   
## Length:7043   
## Class :character   
## Mode :character   
##   
##   
##   
##

## spec\_tbl\_df [7,043 x 21] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ customerID : chr [1:7043] "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...  
## $ gender : chr [1:7043] "Female" "Male" "Male" "Male" ...  
## $ SeniorCitizen : num [1:7043] 0 0 0 0 0 0 0 0 0 0 ...  
## $ Partner : chr [1:7043] "Yes" "No" "No" "No" ...  
## $ Dependents : chr [1:7043] "No" "No" "No" "No" ...  
## $ tenure : num [1:7043] 1 34 2 45 2 8 22 10 28 62 ...  
## $ PhoneService : chr [1:7043] "No" "Yes" "Yes" "No" ...  
## $ MultipleLines : chr [1:7043] "NO PHONE SERVICE" "NO" "NO" "NO PHONE SERVICE" ...  
## $ InternetService : chr [1:7043] "DSL" "DSL" "DSL" "DSL" ...  
## $ OnlineSecurity : chr [1:7043] "NO" "YES" "YES" "YES" ...  
## $ OnlineBackup : chr [1:7043] "YES" "NO" "YES" "NO" ...  
## $ DeviceProtection: chr [1:7043] "NO" "YES" "NO" "YES" ...  
## $ TechSupport : chr [1:7043] "NO" "NO" "NO" "YES" ...  
## $ StreamingTV : chr [1:7043] "NO" "NO" "NO" "NO" ...  
## $ StreamingMovies : chr [1:7043] "NO" "NO" "NO" "NO" ...  
## $ Contract : chr [1:7043] "Month-to-month" "One year" "Month-to-month" "One year" ...  
## $ PaperlessBilling: chr [1:7043] "Yes" "No" "Yes" "No" ...  
## $ PaymentMethod : chr [1:7043] "Electronic check" NA "Mailed check" "Bank transfer (automatic)" ...  
## $ MonthlyCharges : num [1:7043] 29.9 57 53.9 42.3 70.7 ...  
## $ TotalCharges : num [1:7043] 29.9 1889.5 108.2 1840.8 151.7 ...  
## $ Churn : chr [1:7043] "No" "No" "Yes" "No" ...  
## - attr(\*, "problems")= tibble [1 x 5] (S3: tbl\_df/tbl/data.frame)  
## ..$ row : int 4245  
## ..$ col : chr "tenure"  
## ..$ expected: chr "a double"  
## ..$ actual : chr "#NAME?"  
## ..$ file : chr "'C:\\Users\\USER\\Downloads\\Telcom Dataset.csv'"  
## - attr(\*, "spec")=  
## .. cols(  
## .. customerID = col\_character(),  
## .. gender = col\_character(),  
## .. SeniorCitizen = col\_double(),  
## .. Partner = col\_character(),  
## .. Dependents = col\_character(),  
## .. tenure = col\_double(),  
## .. PhoneService = col\_character(),  
## .. MultipleLines = col\_character(),  
## .. InternetService = col\_character(),  
## .. OnlineSecurity = col\_character(),  
## .. OnlineBackup = col\_character(),  
## .. DeviceProtection = col\_character(),  
## .. TechSupport = col\_character(),  
## .. StreamingTV = col\_character(),  
## .. StreamingMovies = col\_character(),  
## .. Contract = col\_character(),  
## .. PaperlessBilling = col\_character(),  
## .. PaymentMethod = col\_character(),  
## .. MonthlyCharges = col\_double(),  
## .. TotalCharges = col\_double(),  
## .. Churn = col\_character()  
## .. )

## 'data.frame': 7043 obs. of 21 variables:  
## $ customerID : Factor w/ 7043 levels "0002-ORFBO","0003-MKNFE",..: 5376 3963 2565 5536 6512 6552 1003 4771 5605 4535 ...  
## $ gender : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...  
## $ SeniorCitizen : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Partner : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1 ...  
## $ Dependents : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 2 ...  
## $ tenure : num 1 34 2 45 2 8 22 10 28 62 ...  
## $ PhoneService : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 1 2 2 ...  
## $ MultipleLines : Factor w/ 3 levels "NO","NO PHONE SERVICE",..: 2 1 1 2 1 NA 3 2 3 1 ...  
## $ InternetService : Factor w/ 3 levels "DSL","FIBER OPTIC",..: 1 1 1 1 2 2 2 1 2 1 ...  
## $ OnlineSecurity : Factor w/ 2 levels "NO","YES": 1 2 2 2 1 1 1 2 1 2 ...  
## $ OnlineBackup : Factor w/ 2 levels "NO","YES": 2 1 2 1 1 1 2 1 1 2 ...  
## $ DeviceProtection: Factor w/ 2 levels "NO","YES": 1 2 1 2 1 2 1 NA 2 1 ...  
## $ TechSupport : Factor w/ 2 levels "NO","YES": 1 1 1 2 1 1 1 1 2 1 ...  
## $ StreamingTV : Factor w/ 2 levels "NO","YES": 1 1 1 1 1 2 NA 1 2 1 ...  
## $ StreamingMovies : Factor w/ 2 levels "NO","YES": 1 1 1 1 1 2 1 1 2 1 ...  
## $ Contract : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1 1 1 2 ...  
## $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...  
## $ PaymentMethod : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 NA 4 1 3 3 2 4 3 1 ...  
## $ MonthlyCharges : num 29.9 57 53.9 42.3 70.7 ...  
## $ TotalCharges : num 29.9 1889.5 108.2 1840.8 151.7 ...  
## $ Churn : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...

During the data inspection, i had to do data type transformation because the data type that was written was wrong. i had to change them to the appropriate data type and we can see that there are three numerical columns while others are factor.

## [1] 0

There are no duplicated columns.

## Rows: 7,043  
## Columns: 20  
## $ gender <fct> Female, Male, Male, Male, Female, Female, Male, Femal~  
## $ SeniorCitizen <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~  
## $ Partner <fct> Yes, No, No, No, No, No, No, No, Yes, No, Yes, No, Ye~  
## $ Dependents <fct> No, No, No, No, No, No, Yes, No, No, Yes, Yes, No, No~  
## $ tenure <dbl> 1, 34, 2, 45, 2, 8, 22, 10, 28, 62, 13, 16, 58, 49, 2~  
## $ PhoneService <fct> No, Yes, Yes, No, Yes, Yes, Yes, No, Yes, Yes, Yes, Y~  
## $ MultipleLines <fct> NO PHONE SERVICE, NO, NO, NO PHONE SERVICE, NO, NA, Y~  
## $ InternetService <fct> DSL, DSL, DSL, DSL, FIBER OPTIC, FIBER OPTIC, FIBER O~  
## $ OnlineSecurity <fct> NO, YES, YES, YES, NO, NO, NO, YES, NO, YES, YES, NO,~  
## $ OnlineBackup <fct> YES, NO, YES, NO, NO, NO, YES, NO, NO, YES, NO, NO, N~  
## $ DeviceProtection <fct> NO, YES, NO, YES, NO, YES, NO, NA, YES, NO, NO, NO, N~  
## $ TechSupport <fct> NO, NO, NO, YES, NO, NO, NO, NO, YES, NO, NO, NO, NO,~  
## $ StreamingTV <fct> NO, NO, NO, NO, NO, YES, NA, NO, YES, NO, NO, NO, YES~  
## $ StreamingMovies <fct> NO, NO, NO, NO, NO, YES, NO, NO, YES, NO, NO, NA, YES~  
## $ Contract <fct> Month-to-month, One year, Month-to-month, One year, M~  
## $ PaperlessBilling <fct> Yes, No, Yes, No, Yes, Yes, Yes, No, Yes, No, Yes, No~  
## $ PaymentMethod <fct> Electronic check, NA, Mailed check, Bank transfer (au~  
## $ MonthlyCharges <dbl> 29.85, 56.95, 53.85, 42.30, 70.70, 99.65, 89.10, 29.7~  
## $ TotalCharges <dbl> 29.85, 1889.50, 108.15, 1840.75, 151.65, 820.50, 1949~  
## $ Churn <fct> No, No, Yes, No, Yes, Yes, No, No, Yes, No, No, No, N~

The customer ID column has been dropped as its not needed for analysis. library dplyr is the library used for dropping columns.

## Warning: package 'psych' was built under R version 4.0.5

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 7039 32.57 28.3 29 31.45 32.62 0 1200 1200 10.15 409.69 0.34

The Average tenure of customer is 32.5 months(sd=28). trimmed average is lesser than 32.5 indicating that are extreme figures in the tenure. some customers didn’t last up to a month while some were reported to have spent 1200 months. library psych is the library used for descriptive statistics.

## vars n mean sd median trimmed mad min max range skew kurtosis  
## X1 1 7041 64.76 30.09 70.35 64.97 35.66 18.25 118.75 100.5 -0.22 -1.26  
## se  
## X1 0.36

The average monthly charge of the customers is 64.76 while some paid as high as 118.75 and some as low as 18.25.

## vars n mean sd median trimmed mad min max range skew  
## X1 1 7031 2283.62 2266.77 1397.65 1970.48 1813.15 18.8 8684.8 8666 0.96  
## kurtosis se  
## X1 -0.23 27.03

The average total charges of customers is 2283.62(sd=2266.77). meanwhile some paid as high as 8684.8 while some as low as 18.8.

DATA CLEANING This is the stage after inspecting where the data is corrected of any abnormality.The data is made cleaned and prepared to be used to train a machine. it involves filling of missing values, data type transformation,detectimg and handling of outliers.

## creating a mode function

### getting the mode of all the categorical columns

## [1] "No"

## [1] 0

## [1] "No"

## [1] 0

## [1] 0

## [1] "NO"

## [1] 0

## [1] "FIBER OPTIC"

## [1] 0

## [1] "NO"

## [1] "NO"

## [1] "NO"

## [1] "NO"

## [1] "NO"

## [1] "Month-to-month"

## [1] "Yes"

## [1] "Electronic check"

## [1] "NO"

## [1] 0

## [1] 0

## [1] 0

## [1] 0

## [1] 0

## [1] 0

## [1] 0

## [1] 0

## [1] 0

## [1] 0

## [1] 0

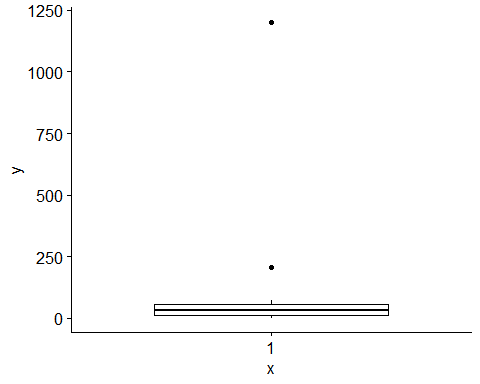
## [1] 0

## gender SeniorCitizen Partner Dependents tenure   
## Female:3488 0:5901 No :3641 No :4933 Min. : 0.00   
## Male :3555 1:1142 Yes:3402 Yes:2110 1st Qu.: 9.00   
## Median : 29.00   
## Mean : 32.57   
## 3rd Qu.: 55.00   
## Max. :1200.00   
## PhoneService MultipleLines InternetService OnlineSecurity  
## No : 682 NO :3392 DSL :2421 NO :5024   
## Yes:6361 NO PHONE SERVICE: 682 FIBER OPTIC:3097 YES:2019   
## YES :2969 NO :1525   
##   
##   
##   
## OnlineBackup DeviceProtection TechSupport StreamingTV StreamingMovies  
## NO :4614 NO :4624 NO :5002 NO :4339 NO :4311   
## YES:2429 YES:2419 YES:2041 YES:2704 YES:2732   
##   
##   
##   
##   
## Contract PaperlessBilling PaymentMethod   
## Month-to-month:3877 No :2872 Bank transfer (automatic):1544   
## One year :1473 Yes:4171 Credit card (automatic) :1520   
## Two year :1693 Electronic check :2370   
## Mailed check :1609   
##   
##   
## MonthlyCharges TotalCharges Churn   
## Min. : 18.25 Min. : 18.8 No :5174   
## 1st Qu.: 35.50 1st Qu.: 402.6 Yes:1869   
## Median : 70.35 Median :1397.7   
## Mean : 64.77 Mean :2282.1   
## 3rd Qu.: 89.85 3rd Qu.:3786.6   
## Max. :118.75 Max. :8684.8

The data is now devoid of missing values. the categorical columns were filled with the most frequently occurring entry while the numerical were filled with their median because the average is greater than the median which suggests the possibility of a positive skewness.

OUTLIER DETECTION AND HANDLING

## Warning: package 'ggpubr' was built under R version 4.0.5



## gender SeniorCitizen Partner Dependents tenure PhoneService MultipleLines  
## 1 Female 0 Yes Yes 1200 Yes YES  
## 2 Female 1 No No 205 Yes NO  
## InternetService OnlineSecurity OnlineBackup DeviceProtection TechSupport  
## 1 FIBER OPTIC YES YES YES NO  
## 2 FIBER OPTIC NO NO NO NO  
## StreamingTV StreamingMovies Contract PaperlessBilling  
## 1 YES YES Two year No  
## 2 NO NO Month-to-month Yes  
## PaymentMethod MonthlyCharges TotalCharges Churn  
## 1 Credit card (automatic) 113.25 7895.15 No  
## 2 Electronic check 69.60 131.65 Yes

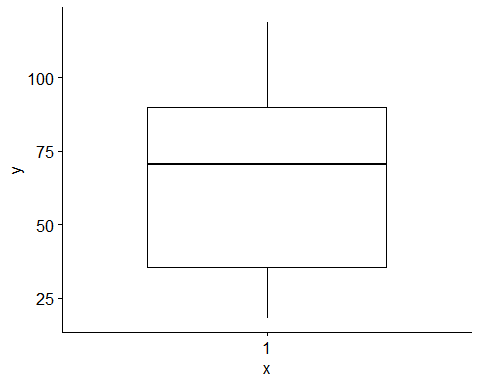
## [1] 2

## [1] 2

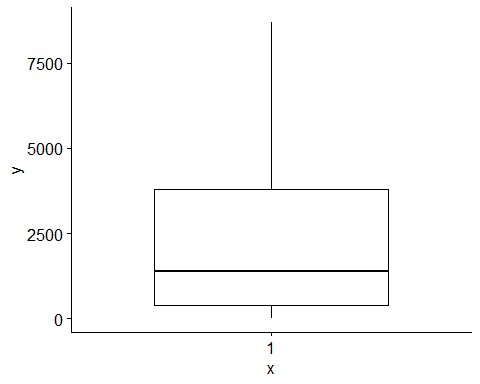
## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 7043 32.38 24.55 29 31.44 32.62 0 72 72 0.24 -1.39 0.29

## [1] 0

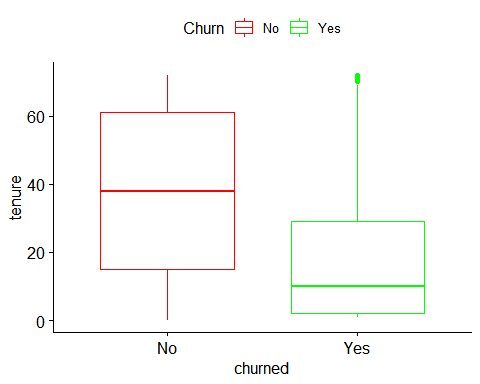
from the box plot above, we can see that the tenure of most of the customers was between 0 months to less than 250 months. however, there are some customers that was reported to have spend around 250 months and even almost 1200 months.In this case, they will be dealt with as outliers. It was dealt with using the inter-quartile range method and converted to missing values that was subsequently replaced with the median as was done above.

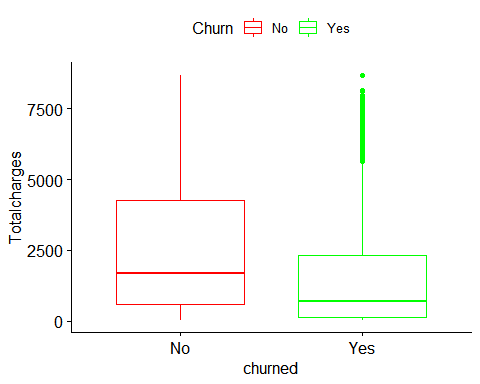


we can see from the box plot above that more than 50% of the customers paid around 75pounds monthly with the minimum being below 25 while some people paid above 100pounds.

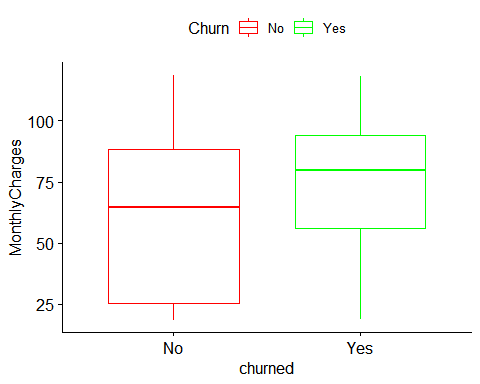
 50% of the customers paid way below 2500 pounds as total charges,while 75% paid around 5000pounds below with some customers paying way above 7500pounds.

EXPLORATORY DATA ANALYSIS is the critical process of performing investigations on data so as to spot anomalies, test hypothesis and to check assumptions with the help of summary statistics and graphical representation.

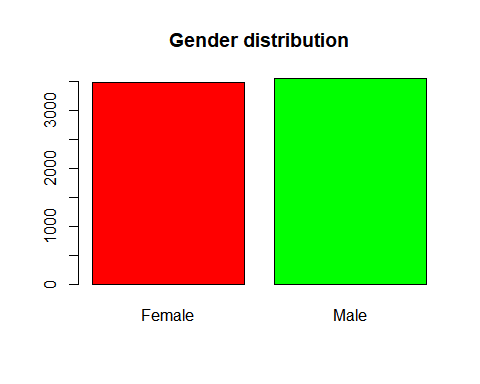
 From the graph above,most of the customers that churned was with the compamy for lesser months than those that didnt churn.



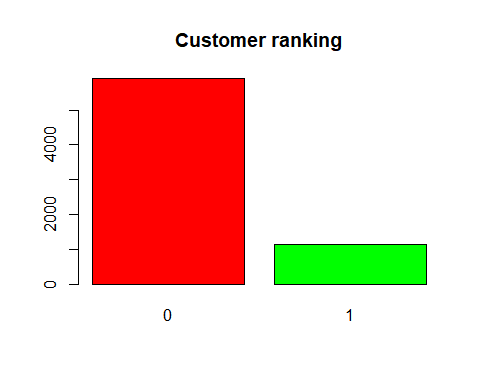
The total amount that was charged most of those customers that were found to have churned is lesser than the customers that stayed. about 75% of the customers that churned paid around 2500pounds although there are some extreme charges of over 7500pounds. meanwhile about 75% of the customers that stayed paid around 5000pounds with some of them also charges a total fee 7500pounds. however, the box plot above shows that most of the churned customers spent very few months as customers.

 the graph above clearly shows that the customers that churned were charged more per month about 50% of them were made to pay around 80pound monthly while 50% of th customers that stayed paid less that 70pound monthly. 25% of those that churned are paying around the same amount with 50% of existeing customer.some of both existing and churned customer were charged above 100 pound monthly. could the hypothesis about price being the cause of churning be true???? could it be that the monthly charges was outrageous, hence churning????

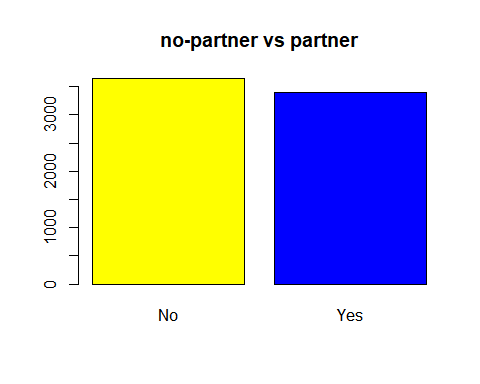
##   
## Female Male   
## 3488 3555

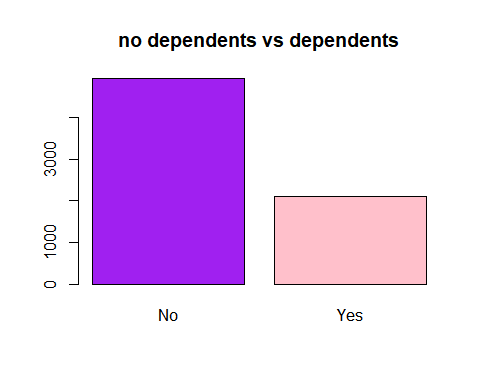
 50.5% of the customers were male while 49.5% were women.

##   
## 0 1   
## 5901 1142

 only 16.2% of the customers were senior citizens, while the remaining were not senior citizens.

##   
## No Yes   
## 3641 3402

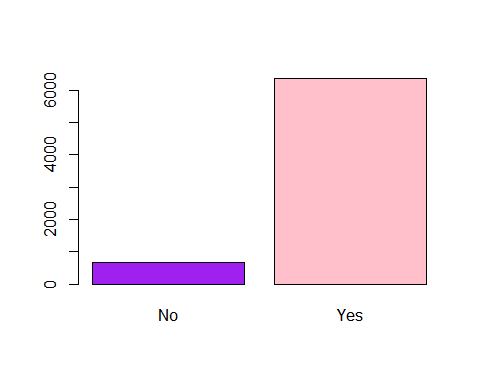
 48.3% of the customers had partners while others didnt have partners.



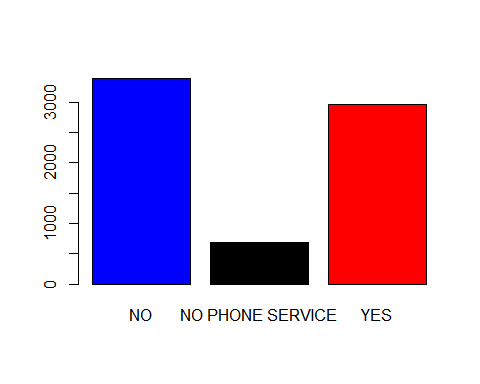
##   
## No Yes   
## 4933 2110

70% of the customers didn’t have dependents while the remaining 30% had dependents.

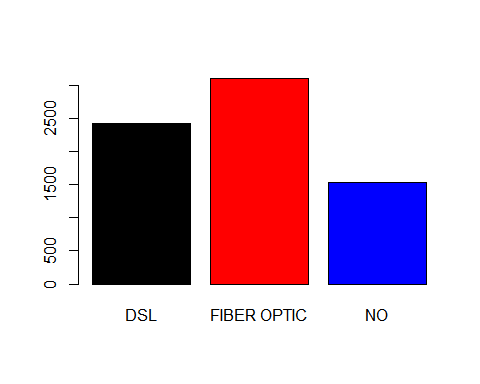
##   
## No Yes   
## 682 6361

 About 90% of the customers used phone service while the remaining didnt use phone service.

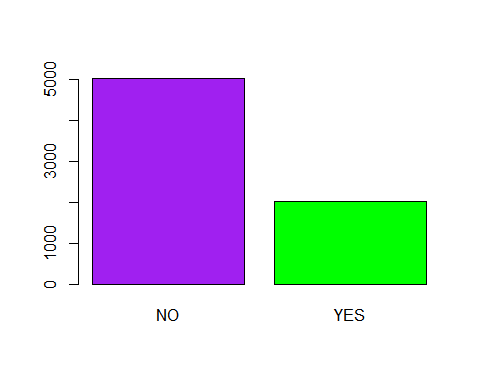
##   
## NO NO PHONE SERVICE YES   
## 3392 682 2969

 About 48% of the customers had just a single line, 42% had multiple lines while the remaining group didnt use phone service.

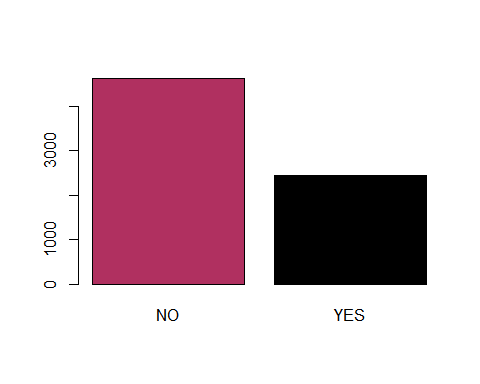
##   
## DSL FIBER OPTIC NO   
## 2421 3097 1525

 21.6% didnt subscribe to an internet service while the remaining 98.4% had access to either fibre optic or DSL internet service.

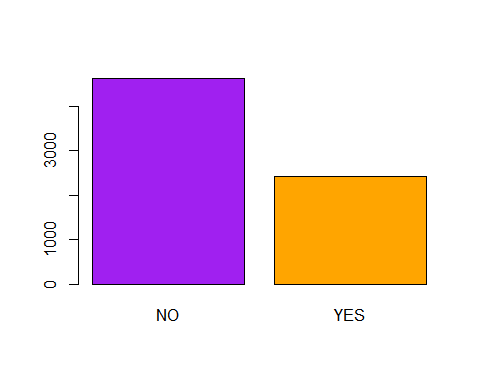
##   
## NO YES   
## 5024 2019

 About 30% subscribed for online security while others didnt.

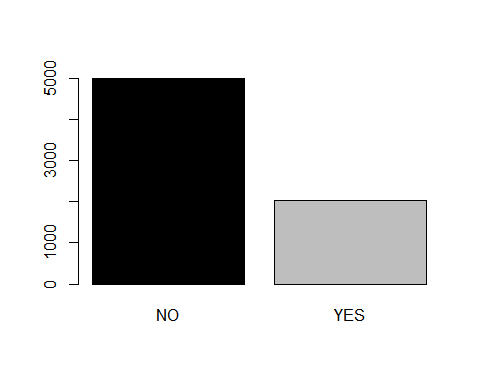
##   
## NO YES   
## 4614 2429

 35% subscribed for online back up while 65% didnt.

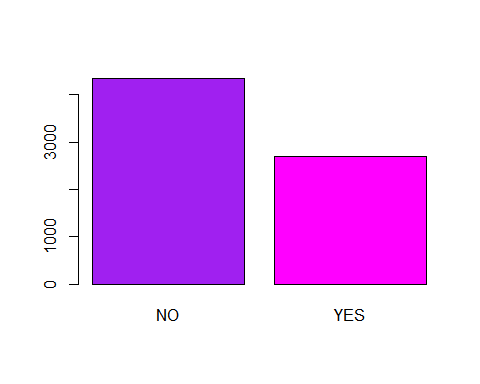
##   
## NO YES   
## 4624 2419

 34.35% of the customers protected their devices with the network provider while the others didnt

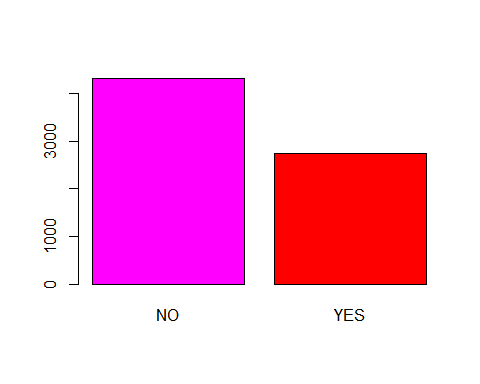
##   
## NO YES   
## 5002 2041

 most of the customers didnt subscribe for the tech support service with only 28.98% subscribing to it.

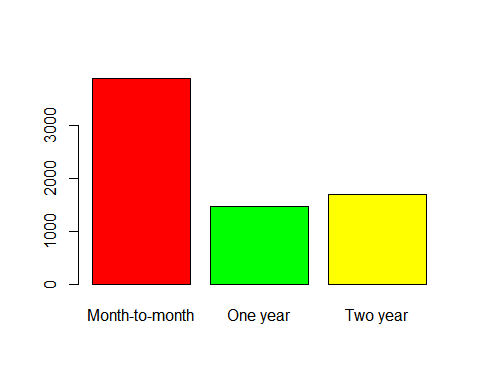
##   
## NO YES   
## 4339 2704

 Only 38.39% would stream Tv programmes online while the others didnt

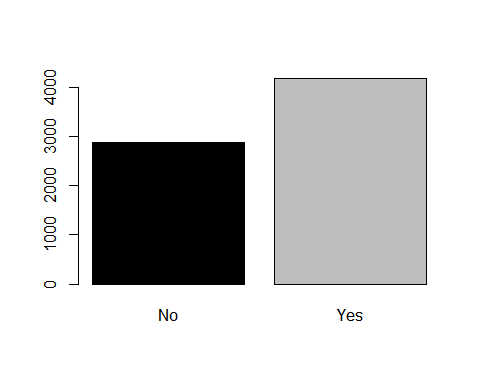
##   
## NO YES   
## 4311 2732

 38.79% streamed movies online while the others didnt.

##   
## Month-to-month One year Two year   
## 3877 1473 1693

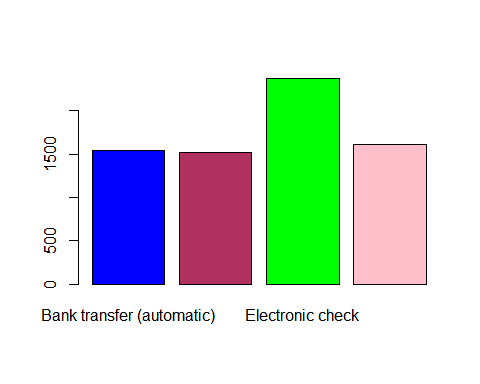
 55% of the customers signed a monthly contract, 20.9% was on one-year contract with the network provider while 24% were on a 2-year contract.

##   
## No Yes   
## 2872 4171



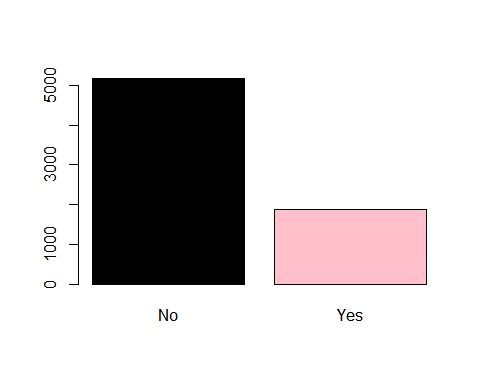
59.22% had paerless billing while others didnt.

##   
## Bank transfer (automatic) Credit card (automatic) Electronic check   
## 1544 1520 2370   
## Mailed check   
## 1609



electronic check was the commonest payment method among the customers with 33.65% users,followed by Mailed check with 22.85% user while the remaining customers used either bank transfer and credit card.

##   
## No Yes   
## 5174 1869

 73.46% 0f the customers are still active customers while the remaining 26.54% has churned.

STATISTICAL ANALYSIS is also the interpretation of data in order to uncober patterns and trends.

##   
## No Yes  
## Female 2549 939  
## Male 2625 930

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: telecom$gender and telecom$Churn  
## X-squared = 0.48408, df = 1, p-value = 0.4866

from the table above, 26.9% of the female churned while 26.2% of the male churned. To establish a relationship between gender and churn status, a chi-square test of independence was run and at 95% confidence interval,the following hypothesis were formed null- ‘there is no statistically significant relationship between the gender of the customers and their churn status’ alternate- ’there is a statistically significant relationship between them.

result; the p-values > 0.05,hence we agree with the null hypothesis. therefore, there is no statiscally significant relationship between the gender ofn the customers and their churn status.

##   
## No Yes  
## 0 4508 1393  
## 1 666 476

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: telecom$SeniorCitizen and telecom$Churn  
## X-squared = 159.43, df = 1, p-value < 2.2e-16

41.7% of senior citizens churned while 23.6% of other citizens churned. we can see that almost half of the senior citizens churned

To establish a relationship between status of the customers and churn status, a chi-square test of independence was run and at 95% confidence interval,the following hypothesis were formed null- ‘there is no statistically significant relationship between the status of the customers and their churn status’ alternate- ’there is a statistically significant relationship between them.

result; the p-values < 0.05,hence we agree with the alternate hypothesis. therefore, there is a statistically significant relationship between the status of the customers and their churn status.

##   
## No Yes  
## No 2441 1200  
## Yes 2733 669

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: telecom$Partner and telecom$Churn  
## X-squared = 158.73, df = 1, p-value < 2.2e-16

64% of people without partner churned while 36% of people with partner churned. we can see that over half of single people churned

To establish a relationship between the marital status of customers and churn status, a chi-square test of independence was run and at 95% confidence interval,the following hypothesis were formed null- ‘there is no statistically significant relationship between the marital status of the customers and their churn status’ alternate- ’there is a statistically significant relationship between them.

result; the p-values < 0.05,hence we agree with the alternate hypothesis. therefore, there is a statistically significant relationship between the marital status and their churn status.

##   
## No Yes  
## No 3390 1543  
## Yes 1784 326

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: telecom$Dependents and telecom$Churn  
## X-squared = 189.13, df = 1, p-value < 2.2e-16

31.3% of people without dependents churned while 15.5% of people with dependents churned.

To establish a relationship between the responsibility status of the customers and churn status, a chi-square test of independence was run and at 95% confidence interval,the following hypothesis were formed null- ‘there is no statistically significant relationship between the responsibility status of the customers and their churn status’ alternate- ’there is a statistically significant relationship between them.

result; the p-values < 0.05,hence we agree with the alternate hypothesis. therefore, there is a statistically significant relationship between the responsibility status of the customers and their churn status.

##   
## No Yes  
## No 512 170  
## Yes 4662 1699

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: telecom$PhoneService and telecom$Churn  
## X-squared = 0.91503, df = 1, p-value = 0.3388

27% of people that subscribed to phone service churned while 25% of people that didn’t use phone service churned.

To establish a relationship between phone service subscription and churn status, a chi-square test of independence was run and at 95% confidence interval,the following hypothesis were formed null- ‘there is no statistically significant relationship between the phone service subscription and churning’ alternate- ’there is a statistically significant relationship between them.

result; the p-values > 0.05,hence we agree with the null hypothesis. therefore, there is no statistically significant relationship between phone service subscription and churning.

##   
## No Yes  
## NO 2541 851  
## NO PHONE SERVICE 512 170  
## YES 2121 848

##   
## Pearson's Chi-squared test  
##   
## data: telecom$MultipleLines and telecom$Churn  
## X-squared = 10.802, df = 2, p-value = 0.004512

28.5% of people with multiple lines churned while 25% of those with a single line churned.

To establish a relationship between the number of lines owned by the customers and churn status, a chi-square test of independence was run and at 95% confidence interval,the following hypothesis were formed null- ‘there is no statistically significant relationship between the number of lines owned by the customers and their churn status’ alternate- ’there is a statistically significant relationship between them.

result; the p-values < 0.05,hence we agree with the alternate hypothesis. therefore, there is a statistically significant relationship between the number of lines owned by the customers and their churn status.

##   
## No Yes  
## DSL 1962 459  
## FIBER OPTIC 1800 1297  
## NO 1412 113

##   
## Pearson's Chi-squared test  
##   
## data: telecom$InternetService and telecom$Churn  
## X-squared = 731.44, df = 2, p-value < 2.2e-16

64% of people without partner churned while 36% of pe churned. we can see that almost half of the senior citizens churned

To establish a relationship between status of the customers and churn status, a chi-square test of independence was run and at 95% confidence interval,the following hypothesis were formed null- ‘there is no statistically significant relationship between the status of the customers and their churn status’ alternate- ’there is a statistically significant relationship between them.

result; the p-values < 0.05,hence we agree with the alternate hypothesis. therefore, there is a statistically significant relationship between the status of the customers and their churn status.

##   
## No Yes  
## NO 3450 1574  
## YES 1724 295

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: telecom$OnlineSecurity and telecom$Churn  
## X-squared = 205.63, df = 1, p-value < 2.2e-16

14.6%% of people that use the online security service churned while 31.3% of those without it churned.

To establish a relationship between online security service and churning, a chi-square test of independence was run and at 95% confidence interval,the following hypothesis were formed null- ‘there is no statistically significant relationship between online security status and churning’ alternate- ’there is a statistically significant relationship between them.

result; the p-values < 0.05,hence we agree with the alternate hypothesis. therefore, there is a statistically significant relationship between online security status and churning.

##   
## No Yes  
## NO 3268 1346  
## YES 1906 523

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: telecom$OnlineBackup and telecom$Churn  
## X-squared = 47.261, df = 1, p-value = 6.214e-12

21.5% of people that subscribed to online backup churned while 29% of non-subscribers of online backup churned.

To establish a relationship between online backup service and churning, a chi-square test of independence was run and at 95% confidence interval,the following hypothesis were formed null- ‘there is no statistically significant relationship between online backup service and churning’ alternate- ’there is a statistically significant relationship between them.

result; the p-values < 0.05,hence we agree with the alternate hypothesis. therefore, there is a statistically significant relationship between online backup service and churning.

##   
## No Yes  
## NO 3300 1324  
## YES 1874 545

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: telecom$DeviceProtection and telecom$Churn  
## X-squared = 30.033, df = 1, p-value = 4.247e-08

28.6% of people without device protection churned while 22.5% of customers with device protection churned.

To establish a relationship between protection of customer’s device and churn status, a chi-square test of independence was run and at 95% confidence interval,the following hypothesis were formed null- ‘there is no statistically significant relationship between protection of customer’s device and churn status’ alternate- ’there is a statistically significant relationship between them.

result; the p-values < 0.05,hence we agree with the alternate hypothesis. therefore, there is a statistically significant relationship between protection of customer’s device and churn status.

##   
## No Yes  
## NO 3443 1559  
## YES 1731 310

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: telecom$TechSupport and telecom$Churn  
## X-squared = 189.03, df = 1, p-value < 2.2e-16

32%% of people without technology support churned while 15% of people with tech support churned.

To establish a relationship between technological support status of customers and their churn status, a chi-square test of independence was run and at 95% confidence interval,the following hypothesis were formed null- ‘there is no statistically significant relationship between technological support status of customers and their churn status’ alternate- ’there is a statistically significant relationship between them.

result; the p-values < 0.05,hence we agree with the alternate hypothesis. therefore, there is a statistically significant relationship between the technological support status of the customers and their churn status.

##   
## No Yes  
## NO 3284 1055  
## YES 1890 814

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: telecom$StreamingTV and telecom$Churn  
## X-squared = 28.343, df = 1, p-value = 1.016e-07

30% of people that stream Tv programmes online churned while 24.3% of those that didnt stream Tv churned.

To establish a relationship between data usage and churn status, a chi-square test of independence was run and at 95% confidence interval,the following hypothesis were formed null- ‘there is no statistically significant relationship between the data usage by customers and their churn status’ alternate- ’there is a statistically significant relationship between them.

result; the p-values < 0.05,hence we agree with the alternate hypothesis. therefore, there is a statistically significant relationship between the data usage of customers and their churn status.

##   
## No Yes  
## NO 3260 1051  
## YES 1914 818

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: telecom$StreamingMovies and telecom$Churn  
## X-squared = 26.251, df = 1, p-value = 2.997e-07

30% of people that streamed movies churned while 24% of customers that didnt stream movies churned.

To establish a relationship between data usage and churn status, a chi-square test of independence was run and at 95% confidence interval,the following hypothesis were formed null- ‘there is no statistically significant relationship between the data usage of customers and their churn status’ alternate- ’there is a statistically significant relationship between them.

result; the p-values < 0.05,hence we agree with the alternate hypothesis. therefore, there is a statistically significant relationship between the data usage of customers and their churn status.

##   
## No Yes  
## Month-to-month 2222 1655  
## One year 1307 166  
## Two year 1645 48

##   
## Pearson's Chi-squared test  
##   
## data: telecom$Contract and telecom$Churn  
## X-squared = 1182.7, df = 2, p-value < 2.2e-16

43% of customers that signed a month-month contract churned, 11% of those in a one-year contract churned while 2.8% of those in a 2-years contract churned. we can see that almost half of those in a monthly contract churned.

To establish a relationship between the contract type of customers and churn status, a chi-square test of independence was run and at 95% confidence interval,the following hypothesis were formed null- ‘there is no statistically significant relationship between the contract type of customers and their churn status’ alternate- ’there is a statistically significant relationship between them.

result; the p-values < 0.05,hence we agree with the alternate hypothesis. therefore, there is a statistically significant relationship between the contract type of customers and their churn status.

##   
## No Yes  
## No 2403 469  
## Yes 2771 1400

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: telecom$PaperlessBilling and telecom$Churn  
## X-squared = 258.28, df = 1, p-value < 2.2e-16

33.5% of customers that were billed on paper churned while 16% of people that werent billed on paper churned.

To establish a relationship between the billing type of customers and churn status, a chi-square test of independence was run and at 95% confidence interval,the following hypothesis were formed null- ‘there is no statistically significant relationship between the billing type of customers and their churn status’ alternate- ’there is a statistically significant relationship between them.

result; the p-values < 0.05,hence we agree with the alternate hypothesis. therefore, there is a statistically significant relationship between the billing type of customers and their churn status.

##   
## No Yes  
## Bank transfer (automatic) 1286 258  
## Credit card (automatic) 1288 232  
## Electronic check 1299 1071  
## Mailed check 1301 308

##   
## Pearson's Chi-squared test  
##   
## data: telecom$PaymentMethod and telecom$Churn  
## X-squared = 643.7, df = 3, p-value < 2.2e-16

45% of people that paid by electronic check were found to have churned, 19% of customers that paid via mailed check while 16% and 15% of those that paid by bank transfer and credit card respectively has churned.

To establish a relationship between the payment mehtod of the customers and churn status, a chi-square test of independence was run and at 95% confidence interval,the following hypothesis were formed null- ‘there is no statistically significant relationship between the payment method of customers and their churn status’ alternate- ’there is a statistically significant relationship between them.

result; the p-values < 0.05,hence we agree with the alternate hypothesis. therefore, there is a statistically significant relationship between the payment method of customers and their churn status.

## Df Sum Sq Mean Sq F value Pr(>F)   
## Churn 1 1.429e+09 1.429e+09 290 <2e-16 \*\*\*  
## Residuals 7041 3.470e+10 4.929e+06   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

To check if there is a statistically significant difference between the average TotalCharges of the customers that churned and current customers. A one-way ANOVA was run and at 95% confidence interval,the following hypothesis were formed null- ‘there is no statistically significant difference between the average total charges of both the churned customers amd the existing customers’ alternate- ’there is a statistically significant difference between them.

result; the p-values < 0.05,hence we agree with the alternate hypothesis. therefore, there is a statistically significant difference between the average totalcharges of the churned and existing customers.

## Df Sum Sq Mean Sq F value Pr(>F)   
## Churn 1 238186 238186 273.2 <2e-16 \*\*\*  
## Residuals 7041 6137525 872   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

To check if there is a statistically significant difference between the average monthly Charges of the customers that churned and current customers. A one-way ANOVA was run and at 95% confidence interval,the following hypothesis were formed null- ‘there is no statistically significant difference between the average monthly charge of both the churned customers and the existing customers’ alternate- ’there is a statistically significant difference between them.

result; the p-values < 0.05,hence we agree with the alternate hypothesis. therefore, there is a statistically significant difference between the average monthly charge of the churned and existing customers.

## Rows: 7,043  
## Columns: 21  
## $ gender <fct> Female, Male, Male, Male, Female, Female, Male, Femal~  
## $ SeniorCitizen <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~  
## $ Partner <fct> Yes, No, No, No, No, No, No, No, Yes, No, Yes, No, Ye~  
## $ Dependents <fct> No, No, No, No, No, No, Yes, No, No, Yes, Yes, No, No~  
## $ tenure <dbl> 1, 34, 2, 45, 2, 8, 22, 10, 28, 62, 13, 16, 58, 49, 2~  
## $ PhoneService <fct> No, Yes, Yes, No, Yes, Yes, Yes, No, Yes, Yes, Yes, Y~  
## $ MultipleLines <fct> NO PHONE SERVICE, NO, NO, NO PHONE SERVICE, NO, NO, Y~  
## $ InternetService <fct> DSL, DSL, DSL, DSL, FIBER OPTIC, FIBER OPTIC, FIBER O~  
## $ OnlineSecurity <fct> NO, YES, YES, YES, NO, NO, NO, YES, NO, YES, YES, NO,~  
## $ OnlineBackup <fct> YES, NO, YES, NO, NO, NO, YES, NO, NO, YES, NO, NO, N~  
## $ DeviceProtection <fct> NO, YES, NO, YES, NO, YES, NO, NO, YES, NO, NO, NO, N~  
## $ TechSupport <fct> NO, NO, NO, YES, NO, NO, NO, NO, YES, NO, NO, NO, NO,~  
## $ StreamingTV <fct> NO, NO, NO, NO, NO, YES, NO, NO, YES, NO, NO, NO, YES~  
## $ StreamingMovies <fct> NO, NO, NO, NO, NO, YES, NO, NO, YES, NO, NO, NO, YES~  
## $ Contract <fct> Month-to-month, One year, Month-to-month, One year, M~  
## $ PaperlessBilling <fct> Yes, No, Yes, No, Yes, Yes, Yes, No, Yes, No, Yes, No~  
## $ PaymentMethod <fct> Electronic check, Electronic check, Mailed check, Ban~  
## $ MonthlyCharges <dbl> 29.85, 56.95, 53.85, 42.30, 70.70, 99.65, 89.10, 29.7~  
## $ TotalCharges <dbl> 29.85, 1889.50, 108.15, 1840.75, 151.65, 820.50, 1949~  
## $ Churn <fct> No, No, Yes, No, Yes, Yes, No, No, Yes, No, No, No, N~  
## $ contractLength <dbl> 1.000000, 33.178227, 2.008357, 43.516548, 2.144979, 8~

A new column(contract length) has been created from Totalcharges and Monthlycharges.The contract length is the exact length of time the customer stayed with the company.

MODEL BUILDING

LABEL ENCODING

## Loading required package: superml

## Warning: package 'superml' was built under R version 4.0.5

## Loading required package: R6

## Warning: package 'R6' was built under R version 4.0.5

Encoding is the process of converting data into a specific code such as numbers.superml is the libarary in R that is responsible for label encoding.

## Warning: package 'caTools' was built under R version 4.0.5

#### splitting the dataset into train and test

The dataset has been splitted into train and test.80% of the dataset would be used for training the machine while the remaining 20% would be used to test the machine. caTools is the library in r that is used to split data into train and test. The test dataset has 21 columns and 1341 observations while the train dataset has the same number of columns with 5702 observations.

##   
## 0 1   
## 4182 1520

 from the barchart above, we can say that 27% of the customers in the training dataset churned while the others are still custoomers.

## Warning: package 'caret' was built under R version 4.0.5

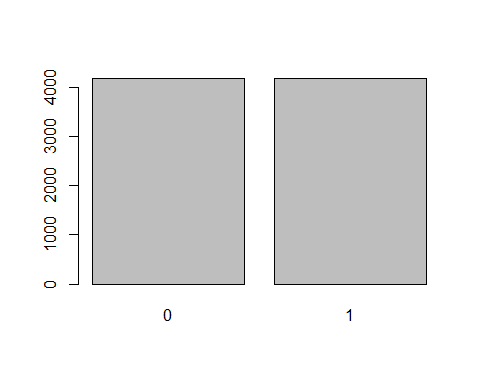
## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

BALANCING THE DATASET by upsampling method

##   
## 0 1   
## 4182 4182

 The training dataset has been upsampled. upsampling is the process of balancing a data by adding to the minority category. here, it has been upsampled and it now has equal number of observations. caret is the library used for balancing dataset aswell as build some models.

Building the logistic model

##   
## Call:  
## glm(formula = Churn ~ ., family = binomial(link = "logit"), data = upsampled\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.1452 -0.7609 0.1665 0.7511 3.1409   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 6.072e-01 1.579e-01 3.845 0.000121 \*\*\*  
## gender -3.602e-02 5.482e-02 -0.657 0.511190   
## SeniorCitizen1 1.403e-01 7.361e-02 1.906 0.056700 .   
## Partner 7.032e-02 6.662e-02 1.056 0.291130   
## Dependents -2.276e-01 7.425e-02 -3.065 0.002178 \*\*   
## tenure -1.043e-02 2.055e-02 -0.507 0.611963   
## PhoneService -1.225e+00 1.276e-01 -9.602 < 2e-16 \*\*\*  
## MultipleLines 1.406e-02 3.496e-02 0.402 0.687516   
## InternetService -1.286e-01 5.233e-02 -2.458 0.013960 \*   
## OnlineSecurity -5.459e-01 7.023e-02 -7.773 7.67e-15 \*\*\*  
## OnlineBackup -3.109e-01 6.721e-02 -4.625 3.74e-06 \*\*\*  
## DeviceProtection -2.805e-01 6.839e-02 -4.101 4.12e-05 \*\*\*  
## TechSupport -5.199e-01 7.157e-02 -7.264 3.75e-13 \*\*\*  
## StreamingTV -1.538e-01 7.491e-02 -2.053 0.040114 \*   
## StreamingMovies -2.635e-02 7.514e-02 -0.351 0.725859   
## Contract -8.340e-01 6.085e-02 -13.707 < 2e-16 \*\*\*  
## PaperlessBilling 2.947e-01 6.167e-02 4.779 1.76e-06 \*\*\*  
## PaymentMethod -8.997e-05 2.965e-02 -0.003 0.997579   
## MonthlyCharges 3.055e-02 2.180e-03 14.013 < 2e-16 \*\*\*  
## TotalCharges 3.421e-04 5.505e-05 6.215 5.12e-10 \*\*\*  
## contractLength -4.748e-02 2.105e-02 -2.255 0.024123 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 11595.0 on 8363 degrees of freedom  
## Residual deviance: 8135.4 on 8343 degrees of freedom  
## AIC: 8177.4  
##   
## Number of Fisher Scoring iterations: 5

## Warning: package 'e1071' was built under R version 4.0.5

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 710 69  
## 1 282 280  
##   
## Accuracy : 0.7383   
## 95% CI : (0.7138, 0.7616)  
## No Information Rate : 0.7397   
## P-Value [Acc > NIR] : 0.5637   
##   
## Kappa : 0.4325   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.7157   
## Specificity : 0.8023   
## Pos Pred Value : 0.9114   
## Neg Pred Value : 0.4982   
## Prevalence : 0.7397   
## Detection Rate : 0.5295   
## Detection Prevalence : 0.5809   
## Balanced Accuracy : 0.7590   
##   
## 'Positive' Class : 0   
##

logistic model is used to model the probability of a certain class or event existing or occuring. from the logistic model above, the accuracy is 74% with a positive predicted value(precision) of 92%.

building a random forest classifier

## Warning: package 'randomForest' was built under R version 4.0.5

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:psych':  
##   
## outlier

## The following object is masked from 'package:dplyr':  
##   
## combine

## The following object is masked from 'package:ggplot2':  
##   
## margin

##   
## Call:  
## randomForest(formula = Churn ~ ., data = upsampled\_train)   
## Type of random forest: classification  
## Number of trees: 500  
## No. of variables tried at each split: 4  
##   
## OOB estimate of error rate: 9.86%  
## Confusion matrix:  
## 0 1 class.error  
## 0 3468 714 0.17073171  
## 1 111 4071 0.02654232

## MeanDecreaseGini  
## gender 80.25678  
## SeniorCitizen 58.34309  
## Partner 67.03506  
## Dependents 68.33690  
## tenure 455.86180  
## PhoneService 21.44030  
## MultipleLines 71.70587  
## InternetService 192.67916  
## OnlineSecurity 85.04048  
## OnlineBackup 67.76713  
## DeviceProtection 58.75632  
## TechSupport 79.69885  
## StreamingTV 55.61957  
## StreamingMovies 64.51060  
## Contract 496.26371  
## PaperlessBilling 87.99829  
## PaymentMethod 175.53707  
## MonthlyCharges 579.05683  
## TotalCharges 513.58223  
## contractLength 587.42617

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 834 123  
## 1 158 226  
##   
## Accuracy : 0.7905   
## 95% CI : (0.7677, 0.812)  
## No Information Rate : 0.7397   
## P-Value [Acc > NIR] : 8.789e-06   
##   
## Kappa : 0.4729   
##   
## Mcnemar's Test P-Value : 0.04253   
##   
## Sensitivity : 0.8407   
## Specificity : 0.6476   
## Pos Pred Value : 0.8715   
## Neg Pred Value : 0.5885   
## Prevalence : 0.7397   
## Detection Rate : 0.6219   
## Detection Prevalence : 0.7136   
## Balanced Accuracy : 0.7441   
##   
## 'Positive' Class : 0   
##

the second model is the random classifer, from the random classifier, we are able to do some feature importance. contract length, monthly charges,total charges,tenure and contract type paid the most role in determining if a customer would churn or not. the random classifer did well with precision of 84% and accuracy of 76.7%

##   
## Call:  
## glm(formula = Churn ~ contractLength + MonthlyCharges + TotalCharges +   
## tenure + Contract + PaymentMethod + InternetService, family = binomial(link = "logit"),   
## data = upsampled\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.0963 -0.8298 0.2086 0.7984 3.1530   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -7.861e-02 1.334e-01 -0.589 0.55561   
## contractLength -4.454e-02 2.076e-02 -2.146 0.03189 \*   
## MonthlyCharges 2.218e-02 1.533e-03 14.471 < 2e-16 \*\*\*  
## TotalCharges 2.462e-04 5.314e-05 4.633 3.61e-06 \*\*\*  
## tenure -6.921e-03 2.027e-02 -0.341 0.73278   
## Contract -1.165e+00 5.695e-02 -20.463 < 2e-16 \*\*\*  
## PaymentMethod -1.077e-02 2.907e-02 -0.370 0.71117   
## InternetService -1.412e-01 4.296e-02 -3.288 0.00101 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 11595.0 on 8363 degrees of freedom  
## Residual deviance: 8482.3 on 8356 degrees of freedom  
## AIC: 8498.3  
##   
## Number of Fisher Scoring iterations: 5

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 666 62  
## 1 326 287  
##   
## Accuracy : 0.7107   
## 95% CI : (0.6856, 0.7348)  
## No Information Rate : 0.7397   
## P-Value [Acc > NIR] : 0.9925   
##   
## Kappa : 0.3965   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.6714   
## Specificity : 0.8223   
## Pos Pred Value : 0.9148   
## Neg Pred Value : 0.4682   
## Prevalence : 0.7397   
## Detection Rate : 0.4966   
## Detection Prevalence : 0.5429   
## Balanced Accuracy : 0.7469   
##   
## 'Positive' Class : 0   
##

A second logistic model was built using some selected features but the performance is lower than the first one. underfitting???

## Length Class Mode   
## call 3 -none- call   
## type 1 -none- character  
## predicted 8364 factor numeric   
## err.rate 1500 -none- numeric   
## confusion 6 -none- numeric   
## votes 16728 matrix numeric   
## oob.times 8364 -none- numeric   
## classes 2 -none- character  
## importance 7 -none- numeric   
## importanceSD 0 -none- NULL   
## localImportance 0 -none- NULL   
## proximity 0 -none- NULL   
## ntree 1 -none- numeric   
## mtry 1 -none- numeric   
## forest 14 -none- list   
## y 8364 factor numeric   
## test 0 -none- NULL   
## inbag 0 -none- NULL   
## terms 3 terms call

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 773 103  
## 1 219 246  
##   
## Accuracy : 0.7599   
## 95% CI : (0.7361, 0.7825)  
## No Information Rate : 0.7397   
## P-Value [Acc > NIR] : 0.04863   
##   
## Kappa : 0.437   
##   
## Mcnemar's Test P-Value : 1.468e-10   
##   
## Sensitivity : 0.7792   
## Specificity : 0.7049   
## Pos Pred Value : 0.8824   
## Neg Pred Value : 0.5290   
## Prevalence : 0.7397   
## Detection Rate : 0.5764   
## Detection Prevalence : 0.6532   
## Balanced Accuracy : 0.7421   
##   
## 'Positive' Class : 0   
##

A second random forest model was also built using same selected features and it performed better than the first one.

MODEL SELECTION

from the report above, we see that the logistic model1 performed better with precision of 91%. hence we will deploy the logistic model.

CONCLUSION

from the Exploratory data analysis ,statistical analysis and the random forest feature importance, I will agree with the hypothesis pf the Lead Data scientist that churning is driven by customer’s price sensitivities. most of the people that churned were people who paid higher monthly charges. Also people that were on month-month contract churned more.

i{r}