**Introduction**

Customer churn occurs when customers or subscribers stop doing business with a company or service, also known as customer attrition. It is also referred as loss of clients or customers. One industry in which churn rates are particularly useful is the telecommunications industry, because most customers have multiple options from which to choose within a geographic location.

Similar concept with [predicting employee turnover](https://towardsdatascience.com/predict-employee-turnover-with-python-da4975588aa3), I am going to predict customer churn using [telecom dataset](https://www.ibm.com/communities/analytics/watson-analytics-blog/guide-to-sample-datasets/). I will introduce **Logistic Regression, Decision Tree, and Random Forest,Naïve bayes**. But this time, I will do all of the above in R. Let’s get started!

### 1)Data Preprocessing

 Each row represents a customer, each column contains that customer’s attributes:

#######setting Working Directory path ######

setwd("Z:/r projects")

Telecom\_churn\_data<-read.csv("Telco-Customer-Churn.csv")

##############required librarys for Project################

library(dplyr)

library(ggplot2)

library(psych)

library(gmodels)

library(randomForest)

library(e1071)

library(caret)

library(rpart)

library(rpart.plot)

library(gplots)

library(nnet)

library(corrplot)

library(car)

> str(Telecom\_churn\_data)

'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 : int 0 0 0 0 0 0 0 0 0 0 ...

$ 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 : int 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 3 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/ 3 levels "No","No internet service",..: 1 3 3 3 1 1 1 3 1 3 ...

$ OnlineBackup : Factor w/ 3 levels "No","No internet service",..: 3 1 3 1 1 1 3 1 1 3 ...

$ DeviceProtection: Factor w/ 3 levels "No","No internet service",..: 1 3 1 3 1 3 1 1 3 1 ...

$ TechSupport : Factor w/ 3 levels "No","No internet service",..: 1 1 1 3 1 1 1 1 3 1 ...

$ StreamingTV : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 3 1 3 1 ...

$ StreamingMovies : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 1 1 3 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 4 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 ...

> glimpse(Telecom\_churn\_data)

Observations: 7,043

Variables: 21

$ customerID <fct> 7590-VHVEG, 5575-GNVDE, 3668-QPYBK, 7795-CFOCW, 9237-HQITU, 9305-CDSKC, 1452-KIOVK, 6713-OKOMC, 7892-POOKP, 6388-TABGU, 976...

$ gender <fct> Female, Male, Male, Male, Female, Female, Male, Female, Female, Male, Male, Male, Male, Male, Male, Female, Female, Male, F...

$ SeniorCitizen <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, ...

$ Partner <fct> Yes, No, No, No, No, No, No, No, Yes, No, Yes, No, Yes, No, No, Yes, No, No, Yes, No, No, Yes, No, Yes, Yes, No, Yes, Yes, ...

$ Dependents <fct> No, No, No, No, No, No, Yes, No, No, Yes, Yes, No, No, No, No, Yes, No, Yes, Yes, No, No, No, No, No, Yes, No, Yes, Yes, No...

$ tenure <int> 1, 34, 2, 45, 2, 8, 22, 10, 28, 62, 13, 16, 58, 49, 25, 69, 52, 71, 10, 21, 1, 12, 1, 58, 49, 30, 47, 1, 72, 17, 71, 2, 27,...

$ PhoneService <fct> No, Yes, Yes, No, Yes, Yes, Yes, No, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, No, Yes, Yes, Yes, Yes, Ye...

$ MultipleLines <fct> No phone service, No, No, No phone service, No, Yes, Yes, No phone service, Yes, No, No, No, Yes, Yes, No, Yes, No, Yes, No...

$ InternetService <fct> DSL, DSL, DSL, DSL, Fiber optic, Fiber optic, Fiber optic, DSL, Fiber optic, DSL, DSL, No, Fiber optic, Fiber optic, Fiber ...

$ OnlineSecurity <fct> No, Yes, Yes, Yes, No, No, No, Yes, No, Yes, Yes, No internet service, No, No, Yes, Yes, No internet service, Yes, No, No, ...

$ OnlineBackup <fct> Yes, No, Yes, No, No, No, Yes, No, No, Yes, No, No internet service, No, Yes, No, Yes, No internet service, No, No, Yes, No...

$ DeviceProtection <fct> No, Yes, No, Yes, No, Yes, No, No, Yes, No, No, No internet service, Yes, Yes, Yes, Yes, No internet service, Yes, Yes, Yes...

$ TechSupport <fct> No, No, No, Yes, No, No, No, No, Yes, No, No, No internet service, No, No, Yes, Yes, No internet service, No, Yes, No, No, ...

$ StreamingTV <fct> No, No, No, No, No, Yes, Yes, No, Yes, No, No, No internet service, Yes, Yes, Yes, Yes, No internet service, Yes, No, No, N...

$ StreamingMovies <fct> No, No, No, No, No, Yes, No, No, Yes, No, No, No internet service, Yes, Yes, Yes, Yes, No internet service, Yes, No, Yes, Y...

$ Contract <fct> Month-to-month, One year, Month-to-month, One year, Month-to-month, Month-to-month, Month-to-month, Month-to-month, Month-t...

$ PaperlessBilling <fct> Yes, No, Yes, No, Yes, Yes, Yes, No, Yes, No, Yes, No, No, Yes, Yes, No, No, No, No, Yes, Yes, No, No, Yes, No, Yes, Yes, N...

$ PaymentMethod <fct> Electronic check, Mailed check, Mailed check, Bank transfer (automatic), Electronic check, Electronic check, Credit card (a...

$ MonthlyCharges <dbl> 29.85, 56.95, 53.85, 42.30, 70.70, 99.65, 89.10, 29.75, 104.80, 56.15, 49.95, 18.95, 100.35, 103.70, 105.50, 113.25, 20.65,...

$ TotalCharges <dbl> 29.85, 1889.50, 108.15, 1840.75, 151.65, 820.50, 1949.40, 301.90, 3046.05, 3487.95, 587.45, 326.80, 5681.10, 5036.30, 2686....

$ Churn <fct> No, No, Yes, No, Yes, Yes, No, No, Yes, No, No, No, No, Yes, No, No, No, No, Yes, No, Yes, No, Yes, No, No, No, Yes, Yes, N...

2)###################**some wrangling and cleanin the data**######################

Features are converted from factores form to numerical form

Telecom\_churn\_data<-Telecom\_churn\_data %>%

mutate(gender\_male=as.numeric(gender=="Male"),

Partner\_yes=as.numeric(Partner=="Yes"),

Dependents\_yes=as.numeric(Dependents=="Yes"),

PhoneService\_yes=as.numeric(PhoneService=="Yes"),

OnlineSecurity\_yes=as.numeric(OnlineSecurity=="Yes"),

OnlineBackup\_yes=as.numeric(OnlineBackup=="Yes"),

DeviceProtection\_yes=as.numeric(DeviceProtection=="Yes"),

TechSupport\_yes=as.numeric(TechSupport=="Yes"),

StreamingTV\_yes=as.numeric(StreamingTV=="Yes"),

StreamingMovies\_yes=as.numeric(StreamingMovies=="Yes"),

PaperlessBilling\_yes=as.numeric(PaperlessBilling=="Yes"),

SeniorCitizen=as.numeric(SeniorCitizen)) %>%

select(-gender,-SeniorCitizen,-Partner,-Dependents,-PhoneService,-MultipleLines,-OnlineSecurity,-DeviceProtection,-TechSupport,-StreamingTV,-StreamingMovies,-PaperlessBilling,-customerID)

Telecom\_churn\_data<-Telecom\_churn\_data %>%

mutate(InternetService\_optic=as.numeric(InternetService == "Fiber optic"),

InternetService\_optic=as.numeric(InternetService == "DSL"))

Telecom\_churn\_data<-Telecom\_churn\_data %>%

mutate(OnlineBackup\_no=as.numeric(OnlineBackup == "No"),

OnlineBackup\_yes=as.numeric(OnlineBackup == "Yes"))

Telecom\_churn\_data<-Telecom\_churn\_data %>%

mutate(Contract\_Month = as.numeric(Contract == "Month-to-month"),

Contract\_two = as.numeric(Contract == "Two year"))

Telecom\_churn\_data<-Telecom\_churn\_data %>%

mutate(payement\_Electronic = as.numeric(PaymentMethod == "Electronic check"),

payement\_Mailed = as.numeric(PaymentMethod == "Mailed check"),

payement\_Bank\_transfer = as.numeric(PaymentMethod == "Bank transfer (automatic)"))

Telecom\_churn\_data<-Telecom\_churn\_data %>%

mutate(Churn=as.numeric(ifelse(Churn == "Yes",1,0))) %>%

select(-InternetService,-OnlineBackup,-Contract,-PaymentMethod)

Telecom\_churn\_data=Telecom\_churn\_data %>%

mutate(tenure\_12=as.numeric(tenure %in% c("0","1","2","4","5","6","7","8","9","10","11","12")),

tenure\_12\_24=as.numeric(tenure > 12 & tenure <= 24 ),

tenure\_60=as.numeric(tenure > 24 & tenure <= 60 )) %>%

select(-tenure)

Remove the columns we do not need for the analysis. after data warngling:-

|  |
| --- |
| > glimpse(Telecom\_churn\_data)  Observations: 7,043  Variables: 24  $ MonthlyCharges <dbl> 29.85, 56.95, 53.85, 42.30, 70.70, 99.65, 89.10, 29.75, 104.80, 56.15, 49.95, 18.95, 100.35, 10...  $ TotalCharges <dbl> 29.85, 1889.50, 108.15, 1840.75, 151.65, 820.50, 1949.40, 301.90, 3046.05, 3487.95, 587.45, 326...  $ Churn <dbl> 0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0,...  $ gender\_male <dbl> 0, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 0, 1, 0, 1, 1, 1, 0, 0, 1,...  $ Partner\_yes <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1,...  $ Dependents\_yes <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0,...  $ PhoneService\_yes <dbl> 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1,...  $ OnlineSecurity\_yes <dbl> 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0,...  $ OnlineBackup\_yes <dbl> 1, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0,...  $ DeviceProtection\_yes <dbl> 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1,...  $ TechSupport\_yes <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0,...  $ StreamingTV\_yes <dbl> 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1,...  $ StreamingMovies\_yes <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1,...  $ PaperlessBilling\_yes <dbl> 1, 0, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1,...  $ InternetService\_optic <dbl> 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 0, 0,...  $ OnlineBackup\_no <dbl> 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1,...  $ Contract\_Month <dbl> 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1,...  $ Contract\_two <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0,...  $ payement\_Electronic <dbl> 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0,...  $ payement\_Mailed <dbl> 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0,...  $ payement\_Bank\_transfer <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,...  $ tenure\_12 <dbl> 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1,...  $ tenure\_12\_24 <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,...  $ tenure\_60 <dbl> 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0,... |
|  |
| |  | | --- | |  | |

We use sapply to check the number if missing values in each columns. We found that there are 11 missing values in “TotalCharges\_partner\_yes” columns. So, let’s remove all rows with missing values.

sapply(Telecom\_churn\_data, function(x) sum(is.na(x)))

MonthlyCharges TotalCharges Churn gender\_male Partner\_yes

0 11 0 0 0

Dependents\_yes PhoneService\_yes OnlineSecurity\_yes OnlineBackup\_yes DeviceProtection\_yes

0 0 0 0 0

TechSupport\_yes StreamingTV\_yes StreamingMovies\_yes PaperlessBilling\_yes InternetService\_optic

0 0 0 0 0

OnlineBackup\_no Contract\_Month Contract\_two payement\_Electronic payement\_Mailed

0 0 0 0 0

payement\_Bank\_transfer tenure\_12 tenure\_12\_24 tenure\_60

0 0 0 0

Code for NA Valus removed from data:-

for(col in names(Telecom\_churn\_data)){

if(sum(is.na(Telecom\_churn\_data[,col]))>0 & !(col %in% Telecom\_churn\_data["TotalCharges"] )){

Telecom\_churn\_data[is.na(Telecom\_churn\_data[,col]),col]=mean(Telecom\_churn\_data[,col],na.rm=T)

}

}

MonthlyCharges TotalCharges Churn gender\_male Partner\_yes

0 0 0 0 0

Dependents\_yes PhoneService\_yes OnlineSecurity\_yes OnlineBackup\_yes DeviceProtection\_yes

0 0 0 0 0

TechSupport\_yes StreamingTV\_yes StreamingMovies\_yes PaperlessBilling\_yes InternetService\_optic

0 0 0 0 0

OnlineBackup\_no Contract\_Month Contract\_two payement\_Electronic payement\_Mailed

0 0 0 0 0

payement\_Bank\_transfer tenure\_12 tenure\_12\_24 tenure\_60

0 0 0 0

### 3)Exploratory data analysis and feature selection

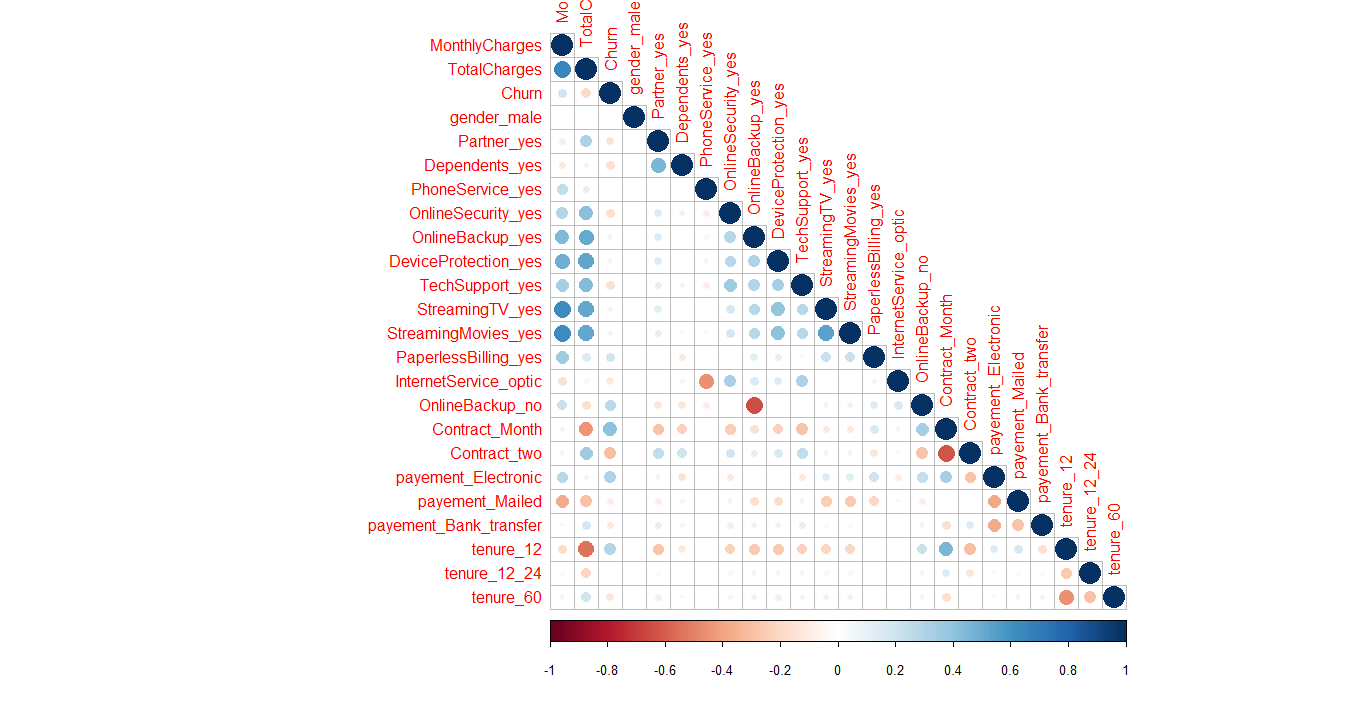
**Correlation between numeric variables**

#######################################

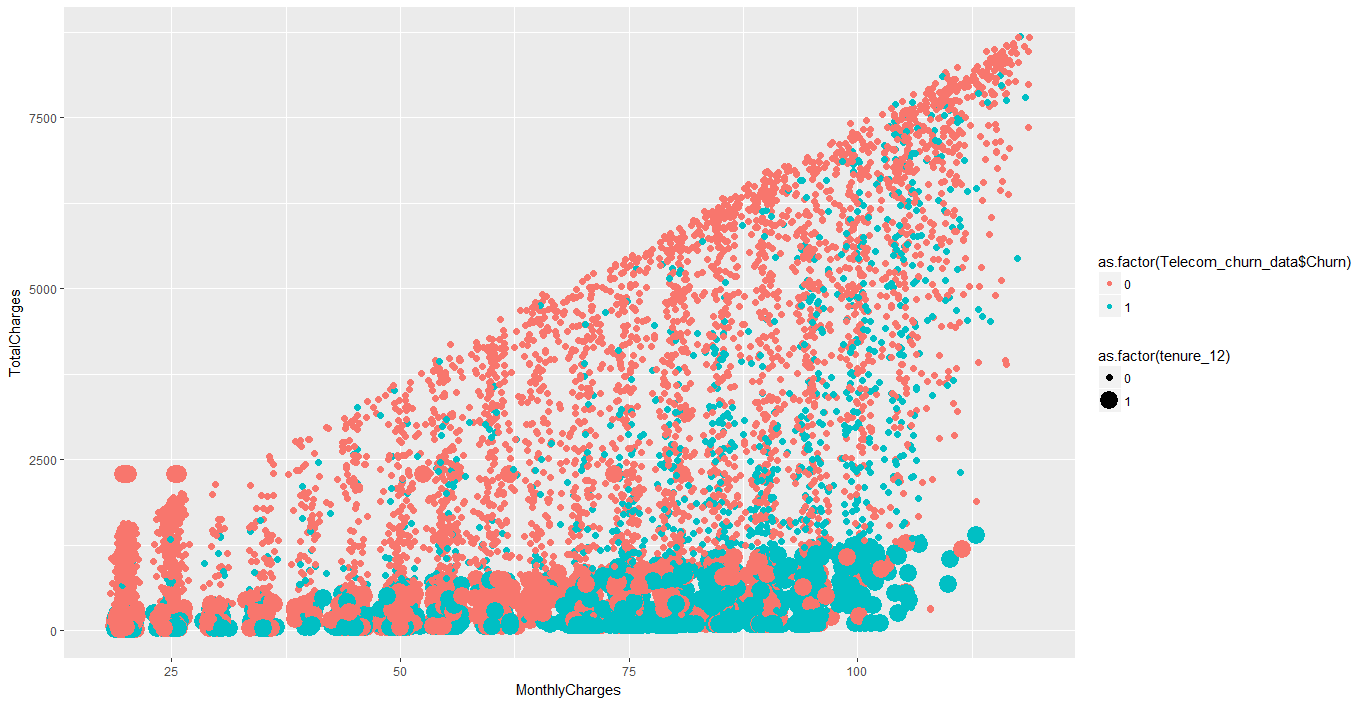
cor1<-cor(Telecom\_churn\_data)

corrplot(cor1,type = "lower")

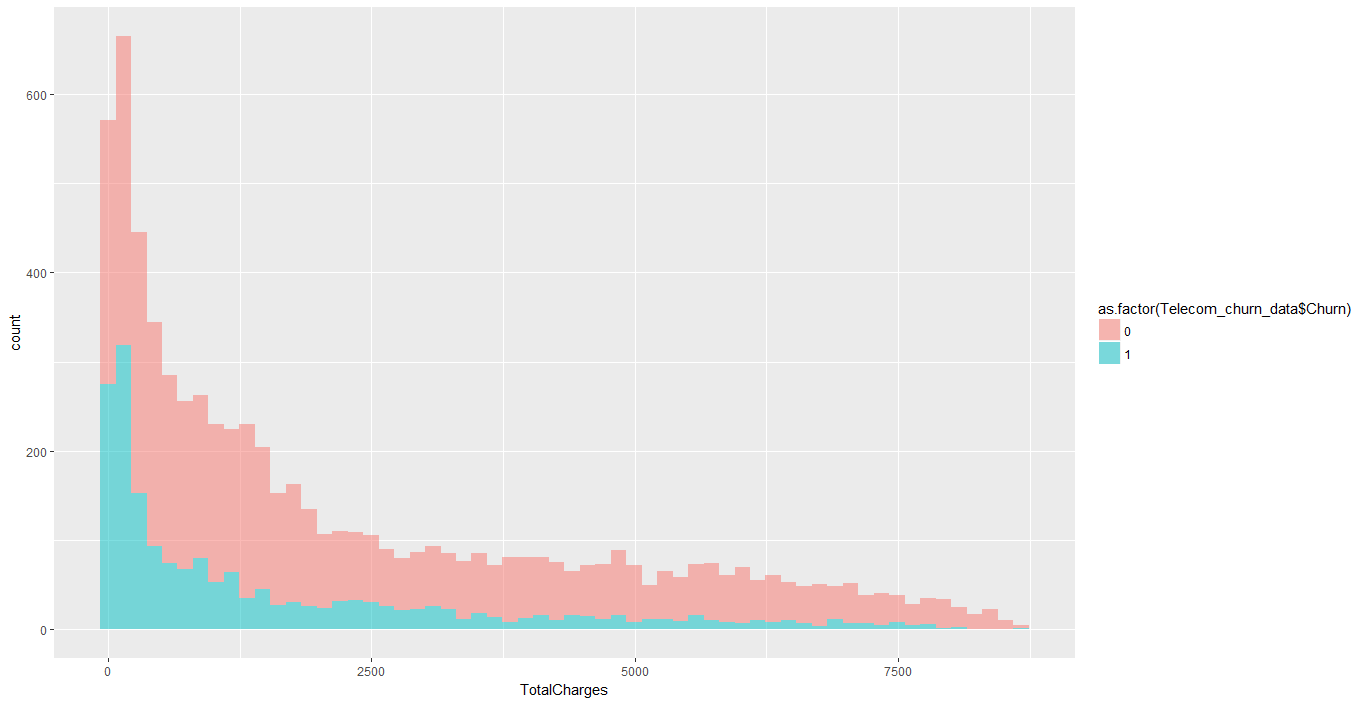
######################################################



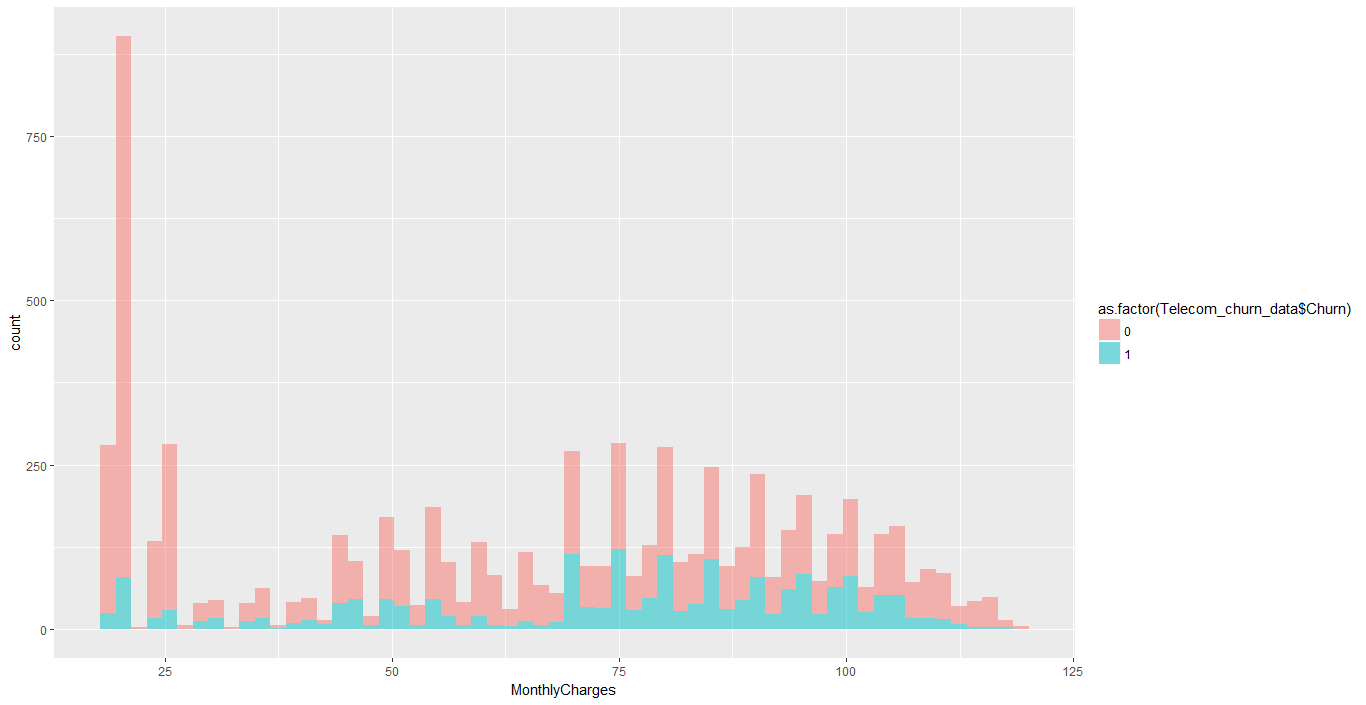
ggplot(Telecom\_churn\_data,aes(y=TotalCharges,x=MonthlyCharges,color=as.factor(Telecom\_churn\_data$Churn),size=as.factor(tenure\_12)))+geom\_point()



ggplot(Telecom\_churn\_data,aes(TotalCharges))+geom\_histogram(aes(fill=as.factor(Telecom\_churn\_data$Churn)),alpha=0.5,bins = 30)



ggplot(Telecom\_churn\_data,aes(MonthlyCharges))+geom\_histogram(aes(fill=as.factor(Telecom\_churn\_data$Churn)),alpha=0.5,bins = 60)



### Logistic Regression

**First, split the data into training and testing sets**:

library(caTools)

set.seed(101)

s=sample.split(Telecom\_churn\_data$Churn,SplitRatio = 0.7)

Telecom\_churn\_data\_train<-subset(Telecom\_churn\_data,s==T)

Telecom\_churn\_data\_test<-subset(Telecom\_churn\_data,s==F)

##################################################################

model\_Telecom\_churn<-glm(Churn~.,data = Telecom\_churn\_data\_train,family = binomial(link = "logit"))

print(summary(model\_Telecom\_churn))

> print(summary(model\_Telecom\_churn))

Call:

glm(formula = Churn ~ ., family = binomial(link = "logit"), data = Telecom\_churn\_data\_train)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.9841 -0.6487 -0.2790 0.6368 3.0893

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.466e+00 2.773e-01 -8.894 < 2e-16 \*\*\*

MonthlyCharges 5.622e-02 1.736e-02 3.239 0.00120 \*\*

TotalCharges -2.031e-04 4.009e-05 -5.066 4.06e-07 \*\*\*

tenure\_12 5.567e-01 1.705e-01 3.265 0.00109 \*\*

tenure\_12\_24 -2.025e-01 1.703e-01 -1.189 0.23442

tenure\_60 -2.430e-01 1.369e-01 -1.775 0.07589 .

gender\_male -1.240e-02 7.838e-02 -0.158 0.87426

Partner\_yes -2.290e-02 9.314e-02 -0.246 0.80576

Dependents\_yes -2.126e-01 1.068e-01 -1.990 0.04660 \*

PhoneService\_yes -1.636e+00 4.025e-01 -4.065 4.80e-05 \*\*\*

OnlineSecurity\_yes -6.822e-01 1.327e-01 -5.141 2.73e-07 \*\*\*

OnlineBackup\_yes -1.256e+00 9.867e-01 -1.273 0.20294

DeviceProtection\_yes -3.484e-01 1.280e-01 -2.722 0.00650 \*\*

TechSupport\_yes -6.197e-01 1.327e-01 -4.668 3.04e-06 \*\*\*

StreamingTV\_yes -2.369e-01 1.979e-01 -1.197 0.23129

StreamingMovies\_yes -2.736e-01 1.973e-01 -1.387 0.16543

PaperlessBilling\_yes 3.899e-01 9.042e-02 4.312 1.62e-05 \*\*\*

InternetService\_optic 3.910e-01 4.562e-01 0.857 0.39137

OnlineBackup\_no -8.637e-01 8.996e-01 -0.960 0.33701

Contract\_Month 8.673e-01 1.271e-01 6.826 8.71e-12 \*\*\*

Contract\_two -1.025e+00 2.206e-01 -4.648 3.36e-06 \*\*\*

payement\_Electronic 3.449e-01 1.148e-01 3.006 0.00265 \*\*

payement\_Mailed -1.489e-02 1.385e-01 -0.108 0.91438

payement\_Bank\_transfer -6.440e-02 1.359e-01 -0.474 0.63564

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 5704.4 on 4929 degrees of freedom

Residual deviance: 4038.9 on 4906 degrees of freedom

AIC: 4086.9

Number of Fisher Scoring iterations: 6

Delete all P Valus whose values are not significant after the final model is :-

model\_Telecom\_churn<-glm(Churn~.-payement\_Mailed-gender\_male-Partner\_yes-payement\_Bank\_transfer-InternetService\_optic-OnlineBackup\_no-StreamingTV\_yes,data = Telecom\_churn\_data\_train,family = binomial(link = "logit"))

Call:

glm(formula = Churn ~ . - payement\_Mailed - gender\_male - Partner\_yes -

payement\_Bank\_transfer - InternetService\_optic - OnlineBackup\_no -

StreamingTV\_yes, family = binomial(link = "logit"), data = Telecom\_churn\_data\_train)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.9818 -0.6538 -0.2795 0.6380 3.0804

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.566e+00 2.210e-01 -11.613 < 2e-16 \*\*\*

MonthlyCharges 3.930e-02 2.702e-03 14.548 < 2e-16 \*\*\*

TotalCharges -1.983e-04 3.842e-05 -5.163 2.43e-07 \*\*\*

tenure\_12 5.503e-01 1.700e-01 3.238 0.001204 \*\*

tenure\_12\_24 -2.035e-01 1.700e-01 -1.197 0.231361

tenure\_60 -2.458e-01 1.367e-01 -1.798 0.072101 .

Dependents\_yes -2.276e-01 9.773e-02 -2.329 0.019857 \*

PhoneService\_yes -1.242e+00 1.554e-01 -7.993 1.32e-15 \*\*\*

OnlineSecurity\_yes -6.066e-01 1.015e-01 -5.974 2.31e-09 \*\*\*

OnlineBackup\_yes -3.124e-01 9.414e-02 -3.318 0.000906 \*\*\*

DeviceProtection\_yes -2.739e-01 9.763e-02 -2.805 0.005030 \*\*

TechSupport\_yes -5.508e-01 1.022e-01 -5.388 7.12e-08 \*\*\*

StreamingMovies\_yes -1.238e-01 1.028e-01 -1.205 0.228311

PaperlessBilling\_yes 3.872e-01 9.016e-02 4.294 1.75e-05 \*\*\*

Contract\_Month 8.780e-01 1.264e-01 6.946 3.75e-12 \*\*\*

Contract\_two -1.021e+00 2.203e-01 -4.635 3.57e-06 \*\*\*

payement\_Electronic 3.700e-01 8.338e-02 4.438 9.08e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 5704.4 on 4929 degrees of freedom

Residual deviance: 4040.8 on 4913 degrees of freedom

AIC: 4074.8

Number of Fisher Scoring iterations: 6

predict\_telecom\_churn<-predict(model\_Telecom\_churn, newdata = Telecom\_churn\_data\_test, type = "response")

library(gmodels)

CrossTable(predict\_telecom\_churn>0.5,Telecom\_churn\_data\_test$Churn,)

**Logistic Regression Confusion Matrix=>**

Cell Contents

|-------------------------|

| N |

| Chi-square contribution |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 2113

| Telecom\_churn\_data\_test$Churn

predict\_telecom\_churn > 0.5 | No | Yes | Row Total |

----------------------------|-----------|-----------|-----------|

FALSE | 1377 | 280 | 1657 |

| 21.016 | 58.142 | |

| 0.831 | 0.169 | 0.784 |

| 0.887 | 0.499 | |

| 0.652 | 0.133 | |

----------------------------|-----------|-----------|-----------|

TRUE | 175 | 281 | 456 |

| 76.369 | 211.273 | |

| 0.384 | 0.616 | 0.216 |

| 0.113 | 0.501 | |

| 0.083 | 0.133 | |

----------------------------|-----------|-----------|-----------|

Column Total | 1552 | 561 | 2113 |

| 0.735 | 0.265 | |

----------------------------|-----------|-----------|-----------|

**Logistic Regression Accuracy** => (1375+284)/2113=0.7851396

### ****Decision Tree****

**Decision Tree visualization**

Telecom\_churn\_data<-Telecom\_churn\_data %>%

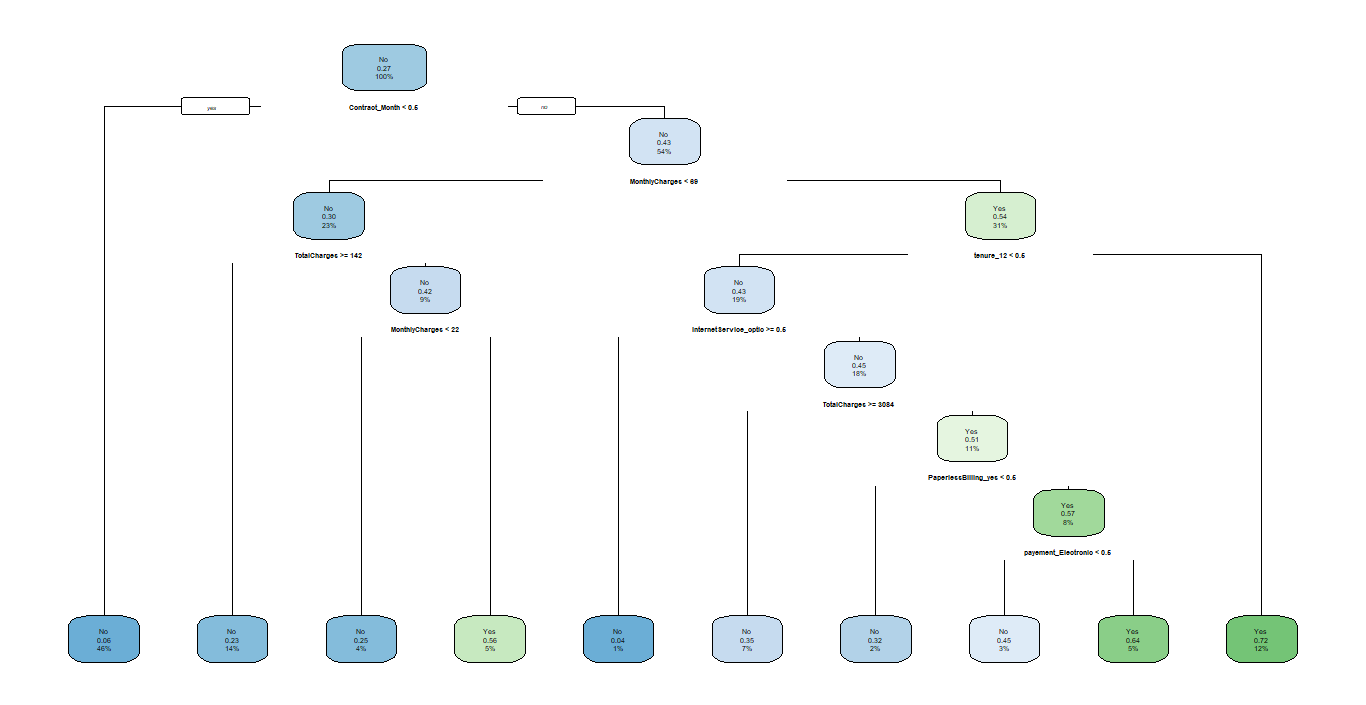
mutate(Churn=as.factor(ifelse(Churn == "1","Yes","No")))

library(rpart)

model\_telecom\_decision\_tree<-rpart(Churn~.,data = Telecom\_churn\_data\_train)

summary(model\_telecom\_decision\_tree)

rpart.plot(model\_telecom\_decision\_tree)



predict\_telecom\_churn\_tree<-predict(model\_telecom\_decision\_tree,newdata = Telecom\_churn\_data\_test,type = "class")

library(caret)

confusionMatrix(Telecom\_churn\_data\_test$Churn,predict\_telecom\_churn\_tree)

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 1368 184

Yes 271 290

Accuracy : 0.7847

95% CI : (0.7665, 0.802)

No Information Rate : 0.7757

P-Value [Acc > NIR] : 0.1674

Kappa : 0.4191

Mcnemar's Test P-Value : 5.536e-05

Sensitivity : 0.8347

Specificity : 0.6118

Pos Pred Value : 0.8814

Neg Pred Value : 0.5169

Prevalence : 0.7757

Detection Rate : 0.6474

Detection Prevalence : 0.7345

Balanced Accuracy : 0.7232

'Positive' Class : No

**Decision Tree Accuracy** is same with logistic regression

The accuracy for Decision Tree has hardly improved. Let’s see if we can do better using Random Forest.

### Random Forest

model\_telcom\_RF<-randomForest(Churn~.,data = Telecom\_churn\_data\_train)

summary(model\_telcom\_RF)

predict\_telecom\_RF<-Predict(model\_telcom\_RF,newdata=Telecom\_churn\_data\_test,type="response")

confusionMatrix(Telecom\_churn\_data\_test$Churn,predict\_telecom\_RF)

**Random Forest Prediction and Confusion Matrix**

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 1354 198

Yes 254 307

Accuracy : 0.7861

95% CI : (0.768, 0.8034)

No Information Rate : 0.761

P-Value [Acc > NIR] : 0.003394

Kappa : 0.4335

Mcnemar's Test P-Value : 0.009682

Sensitivity : 0.8420

Specificity : 0.6079

Pos Pred Value : 0.8724

Neg Pred Value : 0.5472

Prevalence : 0.7610

Detection Rate : 0.6408

Detection Prevalence : 0.7345

Balanced Accuracy : 0.7250

'Positive' Class : No

The Conclusion is Both accuracy and sensitivity are improved ,compare with decision tree and logistic regression

Conclusion:-From the above example, we can see that Logistic Regression, Decision Tree and Random Forest can be used for customer churn analysis for this particular dataset equally fine.

Throughout the analysis, I have learned several important things:

* Features such as tenure\_group, Contract, PaperlessBilling, MonthlyCharges and InternetService appear to play a role in customer churn.
* There does not seem to be a relationship between gender and churn.
* Customers in a month-to-month contract, with PaperlessBilling and are within 12 months tenure, are more likely to churn; On the other hand, customers with one or two year contract, with longer than 12 months tenure, that are not using PaperlessBilling, are less likely to churn.