Telco_Churn_Analysis

Kaustubh

SUPE

Install package

```
install.packages("randomForest")
install.packages("caret")
install.packages("ggplot2")
install.packages("gridExtra")
```

Load libraries

```
library(ggplot2)
library(gridExtra)
library(randomForest)

## randomForest 4.6-14

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

##
## Attaching package: 'randomForest'

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

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

library(caret)

## Loading required package: lattice
```

Load the dataset

```
churn<-read.csv("C:/Users/HP/Documents/Predictive Analysis/Churn.csv")
head(churn)

##   customerID gender SeniorCitizen Partner Dependents tenure PhoneService
## 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
##      MultipleLines InternetService OnlineSecurity OnlineBackup
DeviceProtection
```

```
## 1 No phone service             DSL             No          Yes
No
## 2               No             DSL            Yes           No
Yes
## 3               No             DSL            Yes          Yes
No
## 4 No phone service             DSL            Yes           No
Yes
## 5               No     Fiber optic             No           No
No
## 6              Yes     Fiber optic             No           No
Yes
##   TechSupport StreamingTV StreamingMovies       Contract PaperlessBilling
## 1          No          No              No Month-to-month              Yes
## 2          No          No              No       One year               No
## 3          No          No              No Month-to-month              Yes
## 4         Yes          No              No       One year               No
## 5          No          No              No Month-to-month              Yes
## 6          No         Yes             Yes Month-to-month              Yes
##               PaymentMethod MonthlyCharges TotalCharges Churn
## 1          Electronic check          29.85        29.85    No
## 2              Mailed check          56.95      1889.50    No
## 3              Mailed check          53.85       108.15   Yes
## 4 Bank transfer (automatic)          42.30      1840.75    No
## 5          Electronic check          70.70       151.65   Yes
## 6          Electronic check          99.65       820.50   Yes
```

1. How many customers churn vs no churn?

```
table(churn$Churn)

##
##   No  Yes
## 5174 1869
```

- From the above result, we can see that customers churn = 1869, no churn = 5174

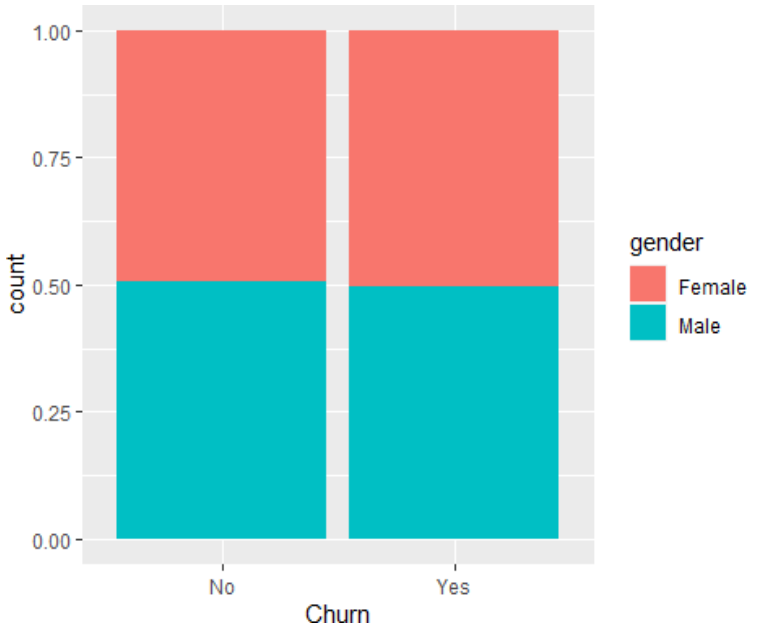
2. Does the gender of customer have an influence on churn?

```
table(churn$gender)

##
## Female   Male
##   3488   3555

plot_gender <- ggplot(data=churn)+geom_bar(mapping = aes(x=Churn,
fill=gender), position="fill")
```

```
plot_gender
```

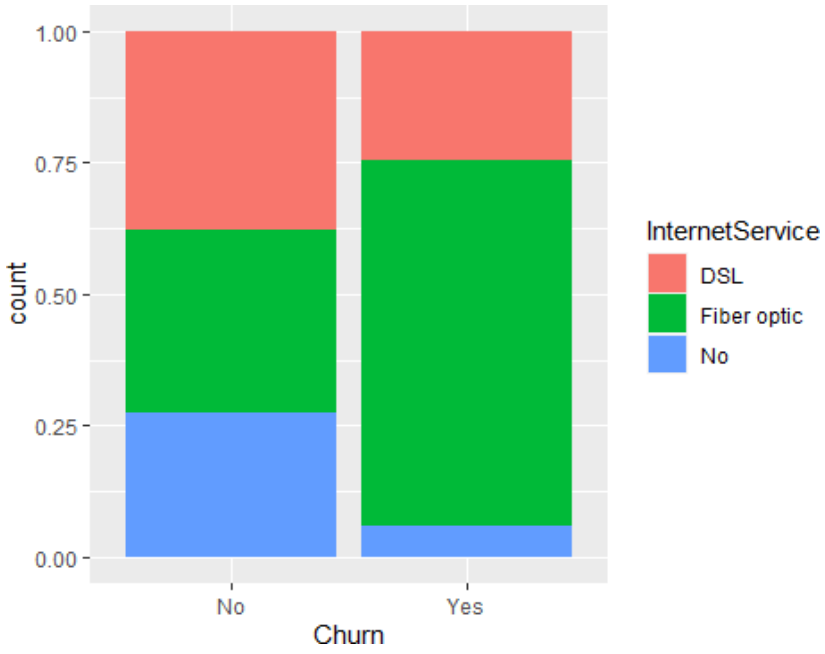


- We can see that gender doesn't have any influence on churn

3. Does the InternetService, TechSupport, .have an influence on Churn ?

InternetService:

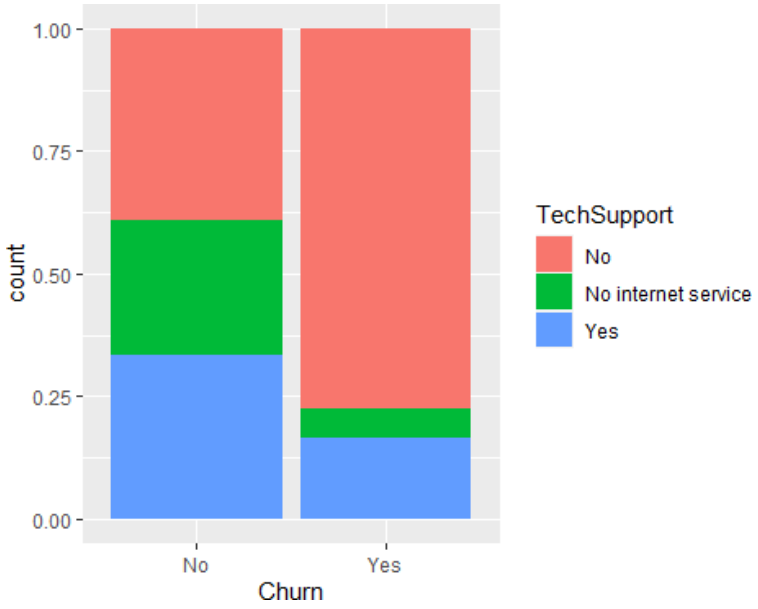
```
plot_InternetService <- ggplot(data=churn)+geom_bar(mapping = aes(x=Churn,
fill=InternetService), position="fill")
plot_InternetService
```



- People using Fiber optic has more tendency to leave the company as it might be some technical issue in Fiber optic Internet service.

TechSupport:

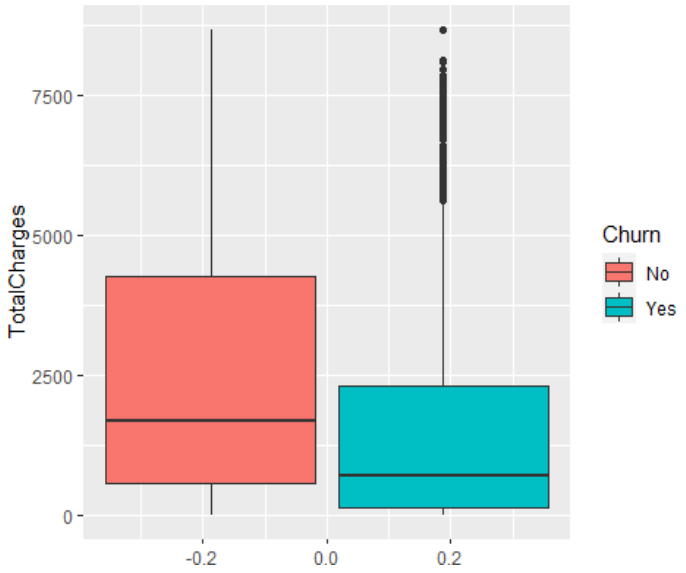
```
plot_TechSupport <- ggplot(data=churn)+geom_bar(mapping = aes(x=Churn, fill=TechSupport), position="fill")
plot_TechSupport
```



- No tech support leads to increase in Churn rate.

4. Does the TotalCharges have an influence on Churn ?

```
ggplot(data=churn)+geom_boxplot(mapping = aes(y= TotalCharges, fill = Churn))
```

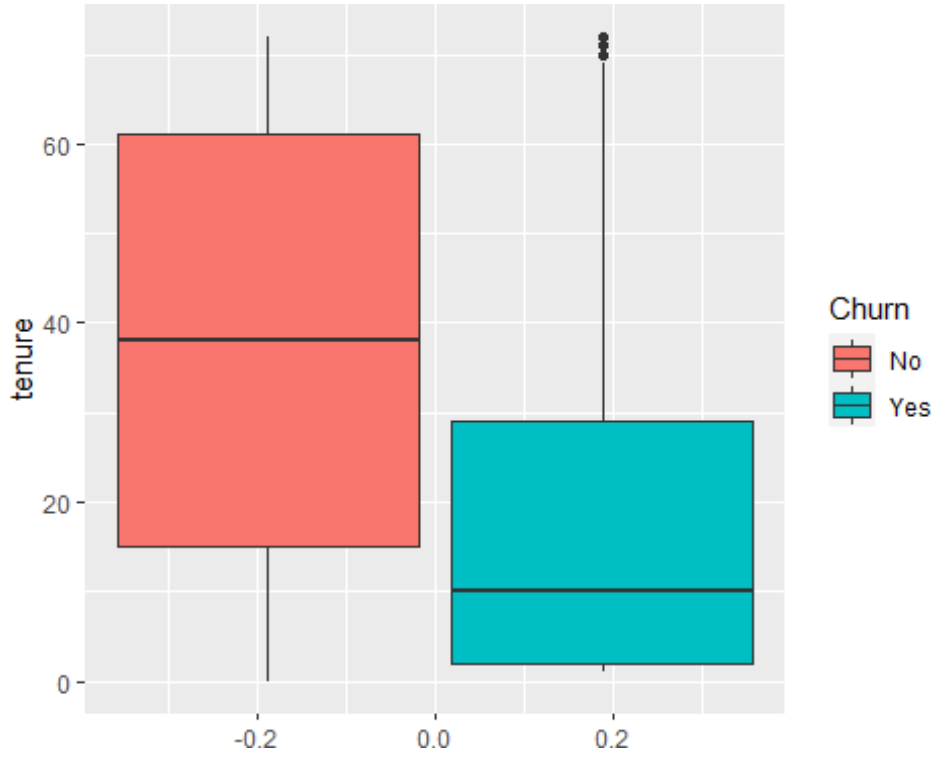


- Looking at the plot, we can say that people left the company are less affected by total charge as median of Churn="Yes" is lower than the one with "No".

5. What is tenure ? How is it linked to Churn ?

- Tenure is the duration of time an employee has worked for in the company.

```
ggplot(data=churn)+geom_boxplot(mapping = aes(y= tenure, fill = Churn))
```



- The more is the tenure, the less chances of customer churns. Here, people having around 10 months of tenure left the company.

6. Which key reasons have probably caused the loss of customers.

To find out this, let's compare graphs of other variables linked to Churn.

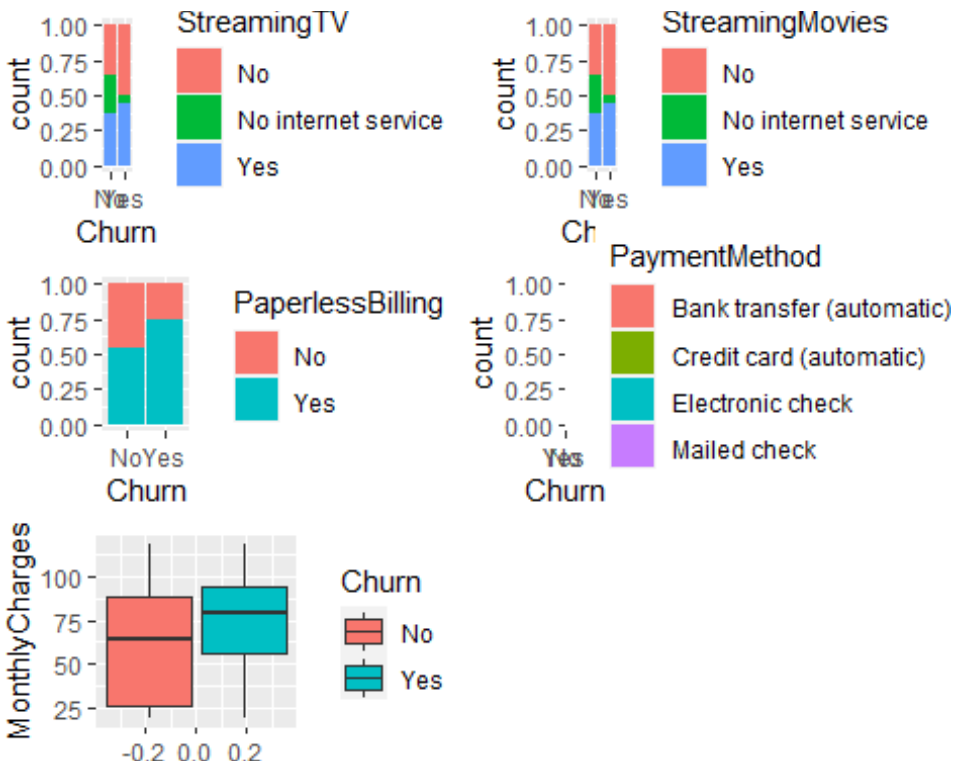
```
plot_SeniorCitizen <- ggplot(data=churn)+geom_bar(mapping = aes(x=Churn,
fill=SeniorCitizen), position="fill")
plot_OnlineSecurity <- ggplot(data=churn)+geom_bar(mapping = aes(x=Churn,
fill=OnlineSecurity), position="fill")
plot_OnlineBackup <- ggplot(data=churn)+geom_bar(mapping = aes(x=Churn,
fill=OnlineBackup), position="fill")
plot_DeviceProtection <- ggplot(data=churn)+geom_bar(mapping = aes(x=Churn,
fill=DeviceProtection), position="fill")

grid.arrange(plot_OnlineSecurity, plot_OnlineBackup, plot_DeviceProtection,
nrow=3)
```

- We can see from the plots that Customers having No Online Security, Online Backup, and Device Protection, left the company.

```
plot_StreamingTV <- ggplot(data=churn)+geom_bar(mapping = aes(x=Churn,
fill=StreamingTV), position="fill")
plot_StreamingMovies <- ggplot(data=churn)+geom_bar(mapping = aes(x=Churn,
fill=StreamingMovies), position="fill")
plot_PaymentMethod <- ggplot(data=churn)+geom_bar(mapping = aes(x=Churn,
fill=PaymentMethod), position="fill")
plot_PaperlessBilling <- ggplot(data=churn)+geom_bar(mapping = aes(x=Churn,
fill=PaperlessBilling), position="fill")
plot_MonthlyCharges <- ggplot(data=churn)+geom_boxplot(mapping = aes(y=
MonthlyCharges, fill = Churn))

grid.arrange(plot_StreamingTV, plot_StreamingMovies, plot_PaperlessBilling,
plot_PaymentMethod, plot_MonthlyCharges, nrow=3)
```



- Paperless billing and Customers with Payment method as electronic check, left the company.

Analyze Dataset

Remove the column id which doesn't influence the prediction

```
churn<-churn[-1]
```

To check NA in features

```
colSums(is.na(churn))
```

```
##           gender    SeniorCitizen          Partner       Dependents
##                0                0                0                0
##           tenure     PhoneService    MultipleLines  InternetService
##                0                0                0                0
##   OnlineSecurity     OnlineBackup DeviceProtection      TechSupport
##                0                0                0                0
##      StreamingTV  StreamingMovies         Contract PaperlessBilling
##                0                0                0                0
##    PaymentMethod   MonthlyCharges     TotalCharges            Churn
##                0                0               11                0
```

- Total Charge has 11 NA which we can ignore.

```
churn <- na.omit(churn)
```

- Rows reduced from 7043 to 7032

Transform Churn into factor

```
churn$Churn<-factor(churn$Churn,levels = c("No","Yes"),labels = c("0","1"))
```

Split the dataset in 2 parts : train and test. Take 80% of lines of dataset churn to create dataset train

```
churn_train<-churn[1:5634,]
churn_test<-churn[5635:7043,]
```

Model fit

```
rf.churn=randomForest(Churn~ . ,data=churn_train)
```

Prediction

```
rfpredict_churn <- predict(rf.churn, churn_test)
confusionMatrix(rfpredict_churn, churn_test$Churn)

## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 919 182
##          1 102 195
##
##                Accuracy : 0.7969
##                  95% CI : (0.7748, 0.8177)
##     No Information Rate : 0.7303
##     P-Value [Acc > NIR] : 4.854e-09
##
##                   Kappa : 0.4473
##
##  Mcnemar's Test P-Value : 2.762e-06
##
##             Sensitivity : 0.9001
##             Specificity : 0.5172
##          Pos Pred Value : 0.8347
##          Neg Pred Value : 0.6566
##              Prevalence : 0.7303
##          Detection Rate : 0.6574
##    Detection Prevalence : 0.7876
##       Balanced Accuracy : 0.7087
##
##        'Positive' Class : 0
##
```

- we achieved 79.26% of accuracy using random forest model.